NAVAL RESEARCH LOGISTICS QUARTERLY

DECEMBER 1969
VOL. 16, NO. 4

OFFICE OF NAVAL RESEARCH

NAVSO P-1278

NAVAL RESEARCH LOGISTICS QUARTERLY

The Naval Research Logistics Quarterly is devoted to the dissemination of scientific information in logistics an will publish research and expository papers, including those in certain areas of mathematics, statistics, and economic relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Information for Contributors is indicated on inside back cover.

The Naval Research Logistics Quarterly is published by the Office of Naval Research in the months of March, June September, and December and can be purchased from the Superintendent of Documents, U.S. Government Printin Office, Washington, D.C. 20402. Subscription Price: $5.50 a year in the U.S. and Canada, $7.00 elsewhere. Cost of individua issues may be obtained from the Superintendent of Documents.

AN OPEN EXPANDING ECONOMY MODEL*

Oskar Morgenstern

Princeton University

and

Gerald L. Thompson

Carnegie-Mellon University

ABSTRACT

The authors extend the generalized von Neumann model they developed (with J. G. Kemeny) in 1956 to an open model by assuming that there are exogeneously determined export and import prices and that any amount can be exported or imported at these prices. The open model is then characterized by means of seven axioms. It is shown, by applying the theory of linear programming, that if four economically reasonable assumptions hold, the open model has at least one solution in which at least one good with positive export price is exported and at least one good with positive import price is imported. It is also shown that, in general, a continuum of expansion rates can be achieved by varying certain control variables. The choice of these expansion rates gives indirectly the choice of a suitable sub-economy and also determines the exports and imports of the economy. Other results and examples are discussed.

1. INTRODUCTION

The exceptional place occupied in the development of mathematical economics by von Neumann's model of an expanding economy [17] is becoming more widely recognized as time progresses. In his brief paper, which is based on a lecture given originally at Princeton University in 1932, and which was published in 1937, the famous minimax theorem, established by John von Neumann in 1928 as the basis of the theory of games, reappears in an essential manner. The appearance of this theorem in two quite separate though related fields is of considerable interest, but the deeper nature of their connection is still largely unexplained as was commented on by von Neumann and Morgenstern [18]; however, the fact that the von Neumann model can be presented and analyzed in terms of matrix game theory and, as will be shown in this paper, also by linear programming procedures (due to the well-known correspondence between them) opens up perhaps one avenue towards explaining this interrelationship.

We shall also find that certain techniques that have been found useful in the linear programming literature enable us to pose economically meaningful conditions on the model and thus extend its scope and usefulness. These conditions enable us further to prove certain new economically meaningful results. We suspect that the same approach will make the proofs of certain turnpike theorems, e.g., those in McKenzie [13], considerably easier. A final advantage of using linear programming is that its theory is now widely known and computational procedures are highly efficient.

---

*The preparation of this paper was supported in part by the Office of Naval Research through the Econometric Research Program at Princeton University, and in part through the Management Sciences Research Group at the Carnegie-Mellon University.

The authors would like to thank Messrs. R. Cornwall, M. Farrell, K. P. Heiss, M. Lovell, and M. Truchon for their valuable comments on earlier drafts of the paper. The responsibility for the final contents rests, of course, with the authors.

It should be remarked that others have suggested the use of linear programming as a tool for analysis of expanding economy models. Three such that have come to our attention are the abstract of a paper by Dorfman [3], Howe's paper [8], and the paper by Haga and Otsuki [6]. In [7], Hamburger, Thompson, and Weil used linear programming to compute expansion rates in the generalized von Neumann model. Gale [5] uses the theory of convex cones in his well-known model, and the connection to linear programming is very close. An interesting proof depending on nonlinear programming has been given by H. Kuhn in [11].

In what follows we assume the reader to be familiar with von Neumann's model and we merely restate some of its principal features: The model is that of a uniformly expanding closed economy for which it is proved that the expansion factor is equal to the interest factor. This fundamental result was obtained from a number of economically plausible assumptions, such as that goods are produced from each other and that processes that are unprofitable will be abandoned. It was also assumed that every good entered somehow, in no matter what small manner, into the production of every other good. This assumption was needed in order to assure that the economy would not break apart into disconnected pieces. From an economic point of view, however, this requirement is at least uncomfortable and should be rejected. This was done by Kemeny-Morgenstern-Thompson [9] in 1956 (henceforth denoted by KMT), who replaced it by the condition that the transformation of inputs into outputs should produce something of value. This is an economically more reasonable requirement.

The consequences of this generalization were that the expanding economy did not break apart, but that it might operate a number of inter-related subeconomies, each one of which having its own expansion rate equal to its interest rate. The number of these rates were finite and placed between determined bounds. (Sub-economies were afterwards studied in great detail by the late K. Förstner in [4].) In addition, outside demand was introduced, and it was shown how its variations influence growth rates and interest rates. Furthermore the generalized model explained the possibility of technological progress in the form of new, more efficient processes as well as the possibility of aggregation and the consequences for the change of expansion and interest rates.

In a subsequent paper [14] in 1967 (henceforth denoted by MT) a further generalization was made by Morgenstern and Thompson introducing savings and investments into the von Neumann model as modified in the 1956 KMT work. Both of these activities could be carried out by private and government agents. These activities may change preferences of this kind, a factor that had not before been studied for this model, and even for others there were no rigorous theorems. The mathematical analysis revealed among other things, that the frequently made assumption that an expanding economy must automatically go to the efficient point is not true in general since, as we have shown, no such point even need to exist. An outside agency, e.g., the government or an ideologically controlled board would have to determine in which of alternative directions the economy would have to expand.

Both of these results—the existence of sub-economies with individual discrete but bounded expansion rates, and the non-existence of efficient points—show the power of mathematical analysis which yields economically meaningful results that could not have been anticipated intuitively or by non-mathematical reasoning. In this there is a profound lesson; others are provided by game theory.

The present study extends our previous work and obtains similarly unexpected results.* We now make the model open, by allowing the economy to be embedded in the "world," which is, as yet,

*The relationship between KMT and MT to other studies of the von Neumann model is fully explained in MT, where it is also demonstrated that the MT model contains all others in specified manners. See also the discussion in Tomasini [16].

not further specified. We shall at a later occasion describe a "world" consisting of several other open-economies which interact with each other in a complicated manner. At present we merely assume that there exists a world market for different goods such that the open expanding economy can import and export freely, provided the prices are right and the balance of payments condition is met. Thus the outside world affects the internal price structure: the internal equilibrium prices have to lie between the import and export prices.

We will find that the fact of openness, when achieved, produces interesting changes: for example, the expansion rates of certain sub-economies, while still being bounded, can now vary continuously rather than remain discrete, and openness is achieved when at least one good is exported and one is imported. It is remarkable that in connecting the economy with the "world" by only one single export and one single import good plus adding control variables this interesting transition to continuity occurs. Normally continuity is the consequence of going over to large numbers, e.g., by introducing asymptotic behavior and the like. It will be seen below how a segment of expansion rates, for which particular expansion rates may vary continuously, is contained in the whole span of permissible expansion rates.

Furthermore, an important issue is resolved which thus far has been a very restrictive feature of the von Neumann model and, as a matter of fact, of all linear models whether considering stationary or expanding systems: This is the condition wherein if more is produced than can be sold profitably the prices of the commodities drop to zero. This is objectionable if the excess production of a commodity can be sold in the export market at a positive price. In the present generalization we allow for unprofitable activities that are maintained at the expense of the profitable ones. Indeed, many production processes need free goods for technological reasons and obviously also government services. It is easy to see that this allows the model to take account of the existence of governments, national defense, and other service institutions or, to explain for example, the maintenance of certain deficit operations as, say, the Vienna Opera in order to promote a profitable Austrian tourist traffic. While this is hardly surprising, the point is once more that there are no rigorous theorems. These are only established in this paper on the firm basis of the modified von Neumann model.

Finally, we note that the rate of expansion of the open economy is partly determined by the government (i.e., the drain the "unprofitable" activities exert on the profitable sector), and partly by the outside world via the world prices to which the economy has to adjust itself. It is highly satisfactory that this result emerges as it is one that might be expected intuitively; however, once more it has to be stated that it is one thing to have an expectation of this sort, but to give a rigorous proof is quite another matter. We note that this result is obtained under the simplest possible circumstances: domestic expansion, presumably dependent on consumption and savings preferences (as analyzed in MT) and influenced via the same processes by the government, plus the price structure offered by a neutral "world." If we drop the neutrality and, instead, consider an economy that is faced with deliberately acting outside open economies the far more complicated situation obtains to which we alluded above. The latter is not the subject matter of the present study.

It is now necessary to make some further brief remarks about the *method* of our analysis. Von Neumann's original paper used functional analysis and he also proved a generalized Brouwer fixed point theorem. The KMT extension applied two-person game theory as a powerful mathematical tool which was possible because game theory rests on the minimax theorem which is basic in the original proof given in KMT. Being able to rely on matrix game theory brought about a vast simplification of the formal treatment of the extended problem. The same method proved useful in MT. This paper

uses linear programming techniques due to the initially mentioned well-known correspondence between a pair of linear programming problems and two-person zero-sum games. Since linear programming is applicable to our present problem which contains the one von Neumann dealt with in 1932 as a special case, it follows that the latter can justifiably be viewed as one of the origins of linear programming methods. This is a by-product of perhaps some interest. Since linear programming is now widely known the fact that we are able to transform the given problem into one of that kind may help to improve the accessibility of the profound thoughts contained in the original work.

We finally mention, or, indeed, emphasize as we did already in KMT that the application of game theory as a mathematical tool does not indicate that the situation described by the model is truly a game situation. For this to be the case deliberate actions of participants, who are sometimes opposed to each other, sometimes are cooperating with each other, have to occur. When we state, for example, that the government sets a certain rate of savings or consumption, this is to be understood as being done in a neutral sense. It is possible to transform the analysis so that there is a real struggle between, say, private consumers and the government as to the allocation of resources and the rate of expansion. Similarly, by using linear programming we do not claim that the underlying situation, as described already in [9] and [14], is necessarily or correctly one in which a central authority decides in a deterministic way about allocation of resources. If anything, this is far from reality and far from the substance of our present discussion of the open economy. Linear programming is a *conceptually* very limited matter: it replaces assumptions of continuous relationships by discontinuous ones and allows for inequalities. This makes linear programming more realistic in applications provided the *basic* condition is met, which is that there must exist a central authority on whose acts *alone* the outcome depends. This must be due to the overriding fact that this authority (person, firm, government) has complete and unchallenged control over *all* variables. The influence of indifferent nature which merely introduces a probabilistic uncertainty may be nullified by simple statistical procedures (cf. [18] p. 10n2). Where there is not complete, central control, i.e., when the outcome depends on several decision makers as in game theory, linear programming does not give the complete answer. It can, however, provide ceteris paribus answers. We need not pursue this matter any further in the present context.

We now proceed to the mathematical analysis and give only such interpretations as have not already been stated in the above text.

2. REVIEW OF THE GENERALIZED VON NEUMANN EXPANDING ECONOMY MODEL

Since the KMT paper is the starting point for the present one, we briefly review it here.

We consider a closed economy that has m processes and n goods. There are two nonnegative $m \times n$ matrices, an input matrix A and an output matrix B. There is an m-component row vector x satisfying

$$x \geqq 0 \text{ and } x f = 1, \tag{1}$$

where f is an m-component column vector of all of whose components are 1's. The components of x measure the normalized intensities with which the processes are operated. Similarly there is an n-component column vector y satisfying

$$y \geqq 0 \text{ and } e y = 1, \tag{2}$$

where e is an n-component row vector all of whose components are 1's. The components of y measure the normalized prices for goods in the economy. Finally there are nonnegative expansion and interest factors α and β.

Given these qualities, we define two others,

$$M_\alpha = B - \alpha A \tag{3}$$

and

$$M_\beta = B - \beta A, \tag{4}$$

and now we can state the axioms of the model. (The meaning of each condition is given in parentheses.)

Axiom (C1)	$x M_\alpha \geqslant 0$	(conservation condition)
Axiom (C2)	$M_\beta y \leqslant 0$	(profitless economy condition)
Axiom (C3)	$x M_\alpha y = 0$	(zero price for overproduced goods)
Axiom (C4)	$x M_\beta y = 0$	(zero intensity for inefficient processes)
Axiom (C5)	$x B y > 0$	(something of value is produced)

It was shown in KMT that in order to have solutions satisfying these axioms, certain assumptions must be made. These assumptions were as follows:

ASSUMPTION A1: $Af > 0$ or equivalently $v(-A) < 0$, where $v(-A)$ is the value of the game $-A$. (Every process must have at least one input.)

ASSUMPTION A2: $eB > 0$ or equivalently $v(B) > 0$, where $v(B)$ is the value of the game B. (Every good can be produced in the economy.)

Given these two assumptions it was shown in KMT and [15] that $\alpha = \beta$ and that there was at least one and at most a finite number of expansion rates for which solutions x and y to the axiom system could be found. Moreover these expansion rates corresponded to sub-economies. Further details and interpretations may be found in KMT.

The method of analysis in that paper involved matrix game theory, but the connection between such games and linear programming is well-known.

3. THE OPEN EXPANDING ECONOMY MODEL

For the new model we shall retain all of the previously defined quantities and add some new ones. What we shall do is to consider the economy in connection with other economies. We shall not in this paper, however, go into the details of the other economies. (We plan to consider that problem in a later paper.) Instead we shall refer merely to the "outside world." The sole effect of the outside world in this model shall be to provide the possibility of exporting and importing various goods in any amounts at stated export and import prices. We treat the export and import prices as exogeneous variables and do not go into their determination here.

The meanings of the quantities A, B, α, β, x, y, e, and f are as previously defined. In addition, we define the new quantities below:

	Dimension	Name	Interpretation
Goods	$1 \times n$	w^+	vector of exports
	$1 \times n$	w^-	vector of imports
	$n \times 1$	p^+	vector of export prices
	$n \times 1$	p^-	vector of import prices
Activities	$m \times 1$	z^+	vector of profits of profitable industries
	$m \times 1$	z^-	vector of negative profits (i.e., losses) of unprofitable industries
	$1 \times m$	t^+	vector of upper bounds on intensities
	$1 \times m$	t^-	vector of lower bounds on intensities.

With these and the previously defined quantities we can now state the axioms of the new model. As before, a brief verbal interpretation of each condition is appended after the axiom.

AXIOM (01)* $xM_\alpha = w^+ - w^-$ (production + imports is sufficient for internal demand + exports).
AXIOM (02)* $M_\beta y = z^+ - z^-$ (value of outputs + negative profits of unprofitable industries = value of inputs + profits of profitable industries).
AXIOM (03)** $w^+p^+ = w^-p^-$ (balance of payments condition).
AXIOM (04)** $t^+z^+ = t^-z^-$ (balance of profits condition).
AXIOM (05) $xBy > 0$ (something of value is produced).
AXIOM (06)*** $t^- \leq x \leq t^+$ (intensity vector is within desired bounds).
AXIOM (07)*** $p^+ \leq y \leq p^-$ (price vector is bounded between the export and the import prices).

We shall drop requirements (1) and (2), which state that x and y are probability vectors. They are not necessary because of the boundedness conditions of Axioms (06) and (07). But we require the following nonnegativity conditions to hold.

$$w^+, w^-, z^+, z^-, \geq 0. \tag{5}$$

The connections between these and the previous axioms are the following: Axiom (01) corresponds to Axiom (C1) and would be identical with it if we also required $w^- = 0$; Axiom (02) corresponds to Axiom (C2) and would be identical with it if we also required $z^+ = 0$; if we multiply the expression in Axiom (01) by y and use Axiom (C1) we get Axiom (03); similarly if we multiply the expression in Axiom (02) by x and use Axiom (C2) we get Axiom (04); finally Axioms (C5) and (05) are identical. The Axioms (06) and (07) are new, but correspond to obviously reasonable economic rules. Note that p^+ and p^-, the export and import prices, are determined exogeneously, i.e., by the outside world, but the variables t^+ and t^-, the upper and lower bounds on intensities of operating processes, are determined endogeneously, i.e., inside the economy.

Even though the variables t^+ and t^- occur as upper and lower bounds and will later appear as constraint parameters in a linear programming problem and also in the objective function of the dual linear programming problem, we should like to emphasize the fact that they are basically *control variables* of the economy. That is, they are selected by a combination of decisions by government agents, managerial agents, and consumers. The exact way in which they are determined will vary from economy to economy. Presumably they should appear also in an economy-wide optimization model whose principal purpose is to determine the correct setting of these control variables, but we shall not discuss that problem here. A final remark is that, since the t variables are controlled by the economy, they may be set so as to interpret the input-output model as permitting capital accumulation by the processes that produce those goods which the economy wants to be produced.

---

*In the first two Axioms we are employing a well-known device from linear programming by defining the vector quantity on the left equal to the difference of two nonnegative slack vectors. These slack vectors later appear in the objective function of a linear programming problem. Charnes and Cooper [1] have named this procedure "goal programming." A well-known result we shall use in the subsequent analysis is the following: at most, only one of the variables w_i^+ or w_i^- can be in the basis in any basic solution. This can be expressed as $w_i^+w_i^- = 0$. The same is true of the pair of variables z_j^+ and z_j^-, that is, $z_j^+z_j^- = 0$. These results follow immediately from linear independence of basis vectors.

**We are indebted to M. Truchon for clarifying the role of Axioms (03) and (04).

***Constraints of this kind have previously been considered by Hamburger, Thompson, and Weil in their constrained game formulation of the model for computing purposes [7].

The balance of payments condition, Axiom (03), is well known in economics. It says that in order to achieve stable long run economic growth, the value of exports must equal the value of imports, when evaluated in terms of the external prices.

The balance of profits condition, Axiom (04), however, is not a standard economic construct. What it says is that in a stable long run growth situation the profits of profitable industries must equal the losses of the unprofitable industries, when internal prices are used and the industries are evaluated at their upper and lower bounds of activity level. The reason for this is that the government can then tax the excess profits of profitable industries and use the tax money to sustain the unprofitable industries as well as finance itself.

We wish to emphasize here that the term "unprofitable industry" does not have a derogatory connotation. It merely designates an industry the value of whose outputs is less than the value of its inputs. The economy may find it desirable to operate such industries, such as the defense industry, the entertainment industry, or various service industries, for reasons other than that of making a profit. The Axiom (04) merely requires that the financial support for such actions is forthcoming.

As in the previous model, solutions to this set of axioms need not exist without making certain assumptions. We shall not use the earlier assumptions (A1) and (A2), but instead use the following new assumptions:

(A3) $$0 \leq p^+ \leq p^-,$$

(A4) $$0 \leq t^- \leq t^+,$$

(A5) $$t^- B p^+ > 0,$$

(A6) $$t^- A p^- > 0.$$

Assumptions (A3) and (A4) are clearly necessary, for without them, we could not possibly find vectors x and y satisfying Axioms (06) and (07).

The meaning of (A5) will become clear later when we establish existence proofs. Briefly stated, however, its meaning is the following: *In order to have an economic solution to the open model which is essentially different from a solution to the closed model, the economy must require the operation with positive intensity of at least one process that produces at least one product having a positive export price.* Stated this way, the requirement is economically significant, since, if this condition were not satisfied, the economy would either not export anything or else export only free goods. And if this were the case, the balance of payments condition would further require that only free goods be imported. Hence if the condition did not hold, the open economy would be, for all practical purposes, closed anyhow.

Similarly the meaning of (A6) is the following: *when the economy evaluates its goods at the highest possible prices, the import prices p^-, then even when operating at its minimum intensities there is a positive demand to import at least one good which has a positive import price.* What this assumption prevents is the possibility of operating at high intensity an industry that requires only inputs that are free, because they have a zero import price and producing from this industry goods that have positive value. Stated in this way the condition is economically meaningful.

Note that Assumptions (A5) and (A6) are essentially restrictions on the control variables t^-. There is still the question as to whether it is *rational* for the economy to satisfy these assumptions. We do not go into that question here.

Note also that if we set $p^+ = 0$ and $p^- = f$ and drop Assumptions (A5) and (A6) we will have a closed economy with control variables. After normalizing prices and activities we will thus have a generalization of the closed model in section 2. We shall make use of this version of the closed model later.

We now turn to the proof that, given Assumptions (A3)–(A6) there always exist solutions to the system of Axioms (01)–(07).

4. EXISTENCE OF SOLUTIONS

As in KMT it is easy to show that the existence of solutions implies immediately that the expansion and interest factor (and rates) are equal.

LEMMA 1: If there are solutions to axioms (01)–(07) then

$$\alpha \leqslant \frac{x\,By}{x\,Ay} \leqslant \beta.$$

PROOF: Assume we have vectors x, y, w^+, w^-, z^+, z^- satisfying these axioms. Then by multiplying the expression in (01) by y and using (03) and (07), we have

$$xM_\alpha y = w^+ y - w^- y \geqslant w^+ p^+ - w^- p^- = 0.$$

Similarly, multiplying the expression in (02) by x and using (04) and (06) gives

$$xM_\beta y = xz^+ - xz^- \leqslant t^+ z^+ - t^- z^- \leqslant 0.$$

Using (3) and (4) and Axiom (05), we then have

$$\alpha \leqslant \frac{x\,By}{x\,Ay} \leqslant \beta.$$

In the rest of this paper we shall show that solutions exist with $\alpha = \beta = xBy/xAy$.

As previously remarked we shall use the theory of linear programming to provide solutions to the model. Consider the linear programming problem:

$$\begin{array}{llll}
(6) & \text{Minimize} & -w^+p^+ + w^-p^-, & \\
(7) & xM_\alpha & -w^+ + w^- = 0, & \\
(8) & -x & & \geqslant -t^+, \\
(9) & x & & \geqslant t^-, \\
(10) & w^+,\ w^- & & \geqslant 0.
\end{array}$$

The derivation of this problem from the axioms is as follows: (9) is Axiom (01); constraints (8) and (9) are formally equivalent to Axiom (06); the term $-w^+p^+$ has been put in the objective function (6) in order to maximize the value of exports; and the term w^-p^- has been put into (6) in order to minimize the value of imports.

The dual problem to this one is derived in a purely formal way using rules described in linear programming texts such as [1, 2]. We assign dual variables y to (7), z^+ to (8), and z^- to (9). Then the dual problem is:

$$(11) \qquad \text{Maximize} \qquad -t^+z^+ + t^-z^-,$$

$$(12) \qquad M_\alpha y \quad -z^+ + z^- = 0,$$

$$(13) \qquad -y \qquad \leqslant -p^+,$$

$$(14) \qquad y \qquad \leqslant p^-,$$

$$(15) \qquad z^+, z^- \qquad \geqslant 0.$$

Again note that (13) and (14) are formally equivalent to Axiom (07), and (12) is Axiom (02).

We now prove that this pair of dual linear programming problems has an optimal solution for every $\alpha \geqslant 0$.

LEMMA 2: If assumptions (A3) and (A4) hold, then both problems (6)–(10) and (11)–(15) have optimal solutions for every $\alpha \geqslant 0$.

PROOF: By the duality theorem of linear programming, one of the problems has an optimal solution if and only if the other one does. We shall show first that both problems have feasible solutions for every $\alpha \geqslant 0$. Choose such an α. Consider first the maximizing problem. By assumption (A3) we can find a y satisfying (13) and (14). Select any such y. Now choose z^+ and z^- as follows: Let $(M_\alpha y)_i$ by the ith component of $M_\alpha y$. Then

$$\begin{array}{lll} \text{if } (M_\alpha y)_i \geqslant 0 & \text{let } z_i^+ = (M_\alpha y)_i & \text{and } z_i^- = 0, \\ \text{if } (M_\alpha y)_i < 0 & \text{let } z_i^+ = 0 & \text{and } z_i^- = -(M_\alpha y)_i. \end{array}$$

The resulting y, z^+, and z^- clearly solve constraints (12)–(15) and provide a feasible solution to the maximizing problem. In an entirely analogous way, one can display a feasible solution to the minimizing problem constraints (7)–(10). Since both problems have feasible solutions, it is well known that they both also have optimal solutions, completing the proof.

Actually the duality theorem of linear programming also provides further properties that connect these two solutions and we proceed now to exploit these properties.

Suppose now, with α fixed, we have solutions x, y, w^+, w^-, z^+, and z^- to these dual problems. By the duality theorem, the optimal values of the pair of problems are equal, i.e.,

$$(16) \qquad -w^+p^+ + w^-p^- = -t^+z^+ + t^-z^-.$$

Also, when multiplying (7) by y, we have

$$(17) \qquad xM_\alpha y - w^+y + w^-y = 0.$$

Similarly, multiplying (12) by x gives

$$(18) \qquad xM_\alpha y - xz^+ + xz^- = 0.$$

Now using, successively, (17), (13), (14), (16), (8), (9), and (18), we obtain the following string of inequalities:

$$(19) \qquad \begin{aligned} -xM_\alpha y &= -w^+y + w^-y, \\ &\leqslant -w^+p^+ + w^-p^-, \\ &= -t^+z^+ + t^-z^-, \\ &\leqslant -xz^+ + xz^-, \\ &= -xM_\alpha y. \end{aligned}$$

Since the first and last terms are identical, every inequality in the string is actually an equality. Using the second equality in (19), we obtain

$$w^+(p^+-y)=w^-(p^--y). \tag{20}$$

Since by (10) w^+ and w^- are nonnegative, and by (13) and (14) both $p^+-y\leqslant 0$ and $p^--y\geqslant 0$, we conclude

$$w^+(p^+-y)=0 \tag{21}$$

$$w^-(p^--y)=0. \tag{22}$$

These results follow because the quantity on the left of (20) is nonpositive while the quantity on the right of (20) is nonnegative.

By similar analysis, using the fourth equality in (19), one can show that

$$(t^--x)z^-=0 \tag{23}$$

$$(t^+-x)z^+=0. \tag{24}$$

So far we have used only assumptions (A3) and (A4). Let us now apply Assumption (A5). Since we have solutions $x\geqslant t^-$ and $y\geqslant p^+$, it follows that

$$x\,By\geqslant t^-Bp^+>0,$$

so that Axiom (05) holds as well.

The only two axioms yet to be satisfied are (03) and (04). For these we must make a special choice of α. First note that on the right of (16) we have the difference between the profits of the unprofitable and the profitable industries and on the left of (16) we have the difference between the value of the imports and the value of the exports. We will now select an α so as to make these two quantities simultaneously satisfy Axioms (03) and (04). To do this we shall need to use Assumptions (A5) and (A6).

LEMMA 3: There exists at least one positive expansion factor α such that

$$-w^+p^++w^-p^-=-t^+z^++t^-z^-=0, \tag{25}$$

i.e., the balance of profits and balance of payments conditions are both satisfied.

PROOF: Equation (16) gives the common value of the two linear programming problems. We shall show that this value is negative for small α and positive for large α. Then, since the value of this linear programming problem is a continuous function of the parameters in its constraints, there is an intermediate value of α yielding (25).

Consider the optimal solution when $\alpha=0$. Then (7) becomes

$$xB-w^++w^-=0. \tag{26}$$

Since $B\geqslant 0$, we have $w^-=0$ since, by linear independence, $w_i^+w_i^-=0$, and

$$-w^+p^+=-xBp^+\leqslant -t^-Bp^+<0, \tag{27}$$

where the last inequality comes from (A5). Hence the optimal solution when $\alpha=0$ yields a negative balance of trade (= balance of profits).

Next consider the optimal solution when α is very large. Equation (12) becomes

$$(B-\alpha A)y-z^{+}+z^{-}=0. \tag{29}$$

Since $0 \leqslant y \leqslant p^{-}$, we have

$$-z^{+}+z^{-}=-(B-\alpha A)y \geqslant -(B-\alpha A)p^{-}. \tag{30}$$

Multiplying on the left by x yields

$$-xz^{+}+xz^{-} \geqslant -x(B-\alpha A)p^{-}. \tag{31}$$

Since $t^{-} \leqslant x \leqslant t^{+}$, we can further deduce the inequality

$$-t^{-}z^{+}+t^{-}z^{-} \geqslant -t^{-}(B-\alpha A)p^{-}=-t^{-}Bp^{-}+\alpha(t^{-}Ap^{-}). \tag{32}$$

Now assumption (A6) says $t^{-}Ap^{-}>0$ so that we can make this quantity arbitrarily large by making α large. The objective function of the dual linear programs is $-t^{+}z^{+}+t^{-}z^{-}$ which differs by the term $-(t^{+}-t^{-})z^{+}$ from (32). Because the lefthand side of (12) decreases with α, z^{+} cannot increase with α, and we see that the objective function of the dual programs must also become arbitrarily large as α becomes large. This completes the proof.

All the previous developments permit us now to state the following existence theorem.

THEOREM 1: If Assumptions (A3)–(A6) hold then there exists at least one α such that there are solutions x, y, w^{+}, w^{-}, z^{+}, and z^{-} satisfying Axioms (01)–(07), and $\beta=\alpha$.

The question now arises as to whether or not solutions to the dual linear programming problems are the only possible solutions to the economic model characterized by Axioms (01)–(07). The answer to this question is positive, as demonstrated by the following theorem.

THEOREM 2: Every solution to Axioms (01)–(07) with $\alpha=\beta$ provides a pair of solutions to the dual linear programming problems (6)–(10) and (11)–(15).

PROOF: Let x, y, w^{+}, w^{-}, z^{+}, and z^{-} be vectors satisfying Axioms (01)–(07) and (5). By Axioms (01), (02), (06), and (07) these vectors are feasible for the dual linear programming problems, that is, they satisfy constraints (7)–(10) and (12)–(15). By Axioms (03) and (04) the objective functions of the two dual problems are both zero, and hence are equal. It follows, therefore, that the above vectors are optimal for the dual programs.

5. PROPERTIES OF SOLUTIONS

Having proved the existence of solutions to the open model given certain economically plausible assumptions, and having shown that these solutions are identical to those of a pair of dual linear programs, we now proceed to characterize and interpret such solutions. We shall especially be interested in economic interpretations.

THEOREM 3: In any solution to the model with $\alpha=\beta$, the following conditions hold:

$$w^{+}y=w^{-}y, \tag{33}$$

$$xz^{+}=xz^{-}. \tag{34}$$

i.e., there is balance of payments and balance of profits with respect to internal prices and activities

PROOF: Equation (19) holds for every set of feasible solutions. At the optimum we also have $xM_\alpha y=0$, and hence (19) implies (33) and (34).

Our next results are obvious consequences of Eqs. (21)–(24). These correspond to the "theorem of the alternative," or the "complementary slackness conditions" in game theory and linear programming, respectively.

THEOREM 4: If $\alpha, x, y, w^+, w^-, z^+, z^-$ are solutions to the open model with $\alpha=\beta$, then the following results hold:

(a) $w_j^+>0$ implies $y_j=p_j^+$
$y_j>p_j^+$ implies $w_j^+=0$

(good j can be exported if, and only if, its internal price equals its export price).

(b) $w_j^->0$ implies $y_j=p_j^-$
$y_j<p_j^-$ implies $w_j=0$

(good j can be imported if, and only if, its internal price equals its import price).

(c) $z_i^+>0$ implies $x_i=t_i^+$
$x_i<t_i^+$ implies $z_i^+=0$

(the ith process can be profitable if, and only if, it is run at maximum intensity).

(d) $z_i^->0$ implies $x_i=t_i^-$
$x_i>t_i^-$ implies $z_i^-=0$

(the ith process can be unprofitable if, and only if, it is run at minimum intensity).

The proofs of these results follow immediately from the nonnegativity conditions and Eqs. (21)–(24).

We now turn to the interpretations of the effects of the parameters p^+, p^-, t^+, and t^- on the resulting expansion rate. Clearly the α produced by the solution depends on these parameters, that is

$$\alpha=\alpha(t^+, t^-, p^+, p^-). \tag{35}$$

The exact form of this functional dependence is, of course, very complicated; however the following facts are evident:

THEOREM 5:

(a) If t^- or p^- are increased or t^+ or p^+ are decreased, then α goes down or stays the same.

(b) If t^- or p^- are decreased or t^+ or p^+ are increased, then α goes up or stays the same.

The proofs of these assertions are obvious consequences of the maximizing and minimizing problems and the effects on them of the right hand sides of their constraints.

We characterize next the range of possible α's. Since p^- and p^+ are given exogenously, we cannot control them and therefore assume they are fixed. We can, however, vary both t^- and t^+ continuously within the ranges determined by (A3), (A4), (A5), and (A6). Hence we shall see that the resulting value of α can vary continuously in an interval. This is in distinct contrast to the closed KMT model for which there was only a discrete set of feasible expansion rates.

THEOREM 6: For the open model there exist maximum and minimum expansion factors α_M and α_m for which there are feasible solutions. Moreover, every α satisfying $\alpha_m \leqslant \alpha \leqslant \alpha_M$ is also a feasible expansion factor for the model.

We shall only sketch the proof of this theorem. As in the proof of Lemma 3, if we consider larger and larger α, a point is reached at which there are no longer any profitable industries. Hence α_M is at most as large as the smallest α for which there are only unprofitable or profitless industries. Similarly, if we consider smaller and smaller $\alpha > 0$, there is some sufficiently small value at which *every* industry is profitable. Thus α_m is at least as large as the largest α for which there are only profitable or profitless industries.

Now let t_m^+, t_m^-, t_M^+, and t_M^- be the control variables that cause the model to achieve expansion rates α_m and α_M. Consider the model with control variables

$$t_k^+ = kt_m^+ + (1-k)t_M^+$$
$$t_k^- = kt_m^- + (1-k)t_M^- .$$

If we consider the dual linear programming problems (6)–(10) and (11)–(15) for these t's we see that the parameters of these problems vary continuously with k. Hence the value of α required to make the values of both programs equal to zero also varies continuously with k. Also, when $k=1$ we have $\alpha = \alpha_m$, and when $k=0$ we have $\alpha = \alpha_M$. Hence intermediate α's can be achieved.

It is interesting to note that the maximum expansion factor in the open model is characterized by the largest α at which there is a profitless *industry* and no profitable industries, whereas the maximum expansion factor in the closed model was the largest α at which there was a profitless *subeconomy* and no profitable subeconomies. It follows, therefore, that the maximum expansion factor in the open model can be larger than that of the corresponding closed model. An example of this is given in the next section.

Another remark is that we can introduce outside demand or foreign aid in the same way that we did in KMT, and thus broaden the range of possible expansion rates. This is a common occurrence in emerging countries.

It should be observed that finding the maximum (or the minimum) expansion rate in the open model is a combinatorial problem of selecting a sub-economy that can operate at the desired rate. For instance, to find the maximum expansion rate one must consider the sub-economies of the closed model that have the largest expansion rates and that also require as input a good having positive import price, and produce as output a good having positive export price. Then each of these must be solved as an open model and the one that yields the highest expansion rate is the one that has to be retained.

6. EXAMPLES

We consider the closed example discussed in [8] and [9] in which there are two goods, labelled essentials and inessentials, and two industries one for producing each good. The matrices are:

$$\begin{array}{l} \text{Essentials industry:} \\ \text{Inessentials industry:} \end{array} \quad A = \begin{array}{c} \begin{array}{cc} E & I \end{array} \\ \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \end{array}, \; B = \begin{array}{c} \begin{array}{cc} E & I \end{array} \\ \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix} \end{array}.$$

Thus if we run just the essentials industry, we can expand with $\alpha = 4$, but if we run both industries, our expansion factor drops to 2.

We make it into an open model by adding the constraints:

$$\begin{aligned} t_1^- &\leqslant x_1 \leqslant t_1^+, \\ t_2^- &\leqslant x_2 \leqslant t_2^+, \\ p_1^+ &\leqslant y_1 \leqslant p_1^-, \\ p_2^+ &\leqslant y_2 \leqslant p_2^-. \end{aligned}$$

We shall assume that $t_1^+ = 1$, and $t_2^- > 0$, that is, that the economy desires to produce some inessentials. We also assume for simplicity that $p_2^- = 1$ and $p_1^+ > 0$. With these assumptions the optimal strategies are:

$$x = (1, t_2^-) \qquad \text{and} \qquad y = \begin{pmatrix} p_1^+ \\ 1 \end{pmatrix}.$$

From this it follows that the expansion rate is

$$\alpha = \frac{4p_1^+ + 2t_2^-}{p_1^+ + p_1^+ t_2^- + t_2^-} = \frac{4(p_1^+/t_2^-) + 2}{(p_1^+/t_2^-) + p_1^+ + 1}.$$

Clearly the smaller t_2^- is, the larger α is, and conversely. It is also clear that, in this case we have $2 < \alpha < 4$, that is the open economy expansion factor lies between the two closed economy expansion factors. For instance, when $p_1^+ = 1/4 = t_2^-$ we have $\alpha = 8/3$.

As a slightly more complicated version of this example we consider the problem with

$$B = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix} \qquad A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}.$$

The closed economy has expansion factors 2, 3, and 4. Suppose we add the constraints

$$\begin{aligned} 0.1 &\leqslant y_1 \leqslant 1 & \qquad 0.3 &\leqslant x_1 \leqslant 0.9 \\ 0.2 &\leqslant y_2 \leqslant 0.9 & \qquad 0.2 &\leqslant x_2 \leqslant 0.8 \\ 0.3 &\leqslant y_3 \leqslant 0.8 & \qquad 0.1 &\leqslant x_3 \leqslant 0.7. \end{aligned}$$

It can then be shown by use of the computing technique described in [7], that the expansion factor is $\alpha = 2.38$ and the solution vectors are:

$$\begin{aligned} x &= (0.9, 0.39, 0.1), \\ w^+ &= (0.3, 0, 0), \\ w^- &= (0, 0, 0.04). \end{aligned}$$

$$y = \begin{pmatrix} 0.1 \\ 0.39 \\ 0.8 \end{pmatrix}, \qquad z^+ = \begin{pmatrix} 0.162 \\ 0 \\ 0 \end{pmatrix}, \qquad z^- = \begin{pmatrix} 0 \\ 0 \\ 1.46 \end{pmatrix}.$$

Note that x_2 and y_2 lie in between their respective bounds.

Our final example is a modification of one proposed by M. Truchon. The input-output matrices are:

$$B = \begin{pmatrix} 4 & 0 \\ 0 & 3 \end{pmatrix} \qquad A = \begin{pmatrix} 1 & 2 \\ 1 & 0 \end{pmatrix}.$$

As a closed model its expansion factor is 1.81. Now make it into an open model with constraints

$$0.5 \leq x_1 \leq 1 \qquad 0.5 \leq y_1 \leq 0.5$$
$$0.5 \leq x_2 \leq 1 \qquad 0.5 \leq y_2 \leq 0.5.$$

Now its expansion factor is $\alpha = 2$ and its solution vectors are

$$x = (0.5, 1),$$
$$w^+ = (0, 1),$$
$$w^- = (1, 0).$$
$$y = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}, \qquad z^+ = \begin{pmatrix} 0 \\ 0.5 \end{pmatrix}, \qquad z^- = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

This example shows that the open model may have an expansion factor greater than that of the closed model provided there is an industry that remains profitable at the higher rate.

BIBLIOGRAPHY

[1] Charnes, A. and W. W. Cooper, *Management Models and Industrial Applications of Linear Programming* (Wiley, New York, 1961), two volumes.
[2] Dantzig, G. B., *Linear Programming and Extensions* (Princeton University Press, Princeton, N.J., 1963).
[3] Dorfman, R., Abstract on pages 289–90 of [12].
[4] Förstner, K., "Wirtschaftliches Wachstum bei vollständiger Konkurrenz und linearer Technologie" in *Methods of Operations Research* (Editor, R. Henn) *IV*, 1–258 (1967).
[5] Gale, D., "The closed linear model of production" in *Linear Inequalities and Related Systems* (Editors, H. W. Kuhn and A. W. Tucker) (Princeton University Press, Princeton, N.J., 1956), pp. 285–303.
[6] Haga, H. and M. Otsuki, "On a Generalized von Neumann Model," *International Economic Review*, *6*, 115–23 (1965).
[7] Hamburger, M. J., G. L. Thompson, and R. L. Weil, Jr., "Computation of Expansion Rates for the Generalized von Neumann Model of an Expanding Economy," *Econometrica*, *35*, 542–547 (1967).
[8] Howe, C. W., "An Alternative Proof of the Existence of General Equilibrium in a von Neumann Model," *Econometrica*, *28*, 635–639 (1960).
[9] Kemeny, J. G., O. Morgenstern, and G. L. Thompson, "A Generalization of the von Neumann Model of an Expanding Economy," *Econometrica*, *24*, 115–135 (1956).
[10] Kemeny, J. G., J. L. Snell, and G. L. Thompson, *Introduction to Finite Mathematics* (Prentice-Hall Publishing, Englewood Cliffs, N.J., 1966), Second Edition.
[11] Kuhn, H. W., "Lectures on Mathematical Economics," in *Mathematics of the Decision Sciences* (Editors, G. B. Dantzig and A. F. Veinott, Jr.), Amer. Math. Soc., *12*, 44–84 (1968).
[12] Malinvaud, E. and M. Bacharach (Editors), *Activity Analysis in the Theory of Growth and Planning* (Macmillan, 1967).
[13] McKenzie, L., "Maximal Paths in the von Neumann Model," in [12], pp. 43–63.
[14] Morgenstern, O. and G. L. Thompson, "Private and Public Consumption and Savings in the von Neumann Model of an Expanding Economy," *Kyklos*, *20*, 387–409 (1967).
[15] Thompson, G. L., "On the Solution of a Game-Theoretic Problem" in *Linear Inequalities and Related Systems* (Editors, H. W. Kuhn and A. W. Tucker) (Princeton University Press, Princeton, N.J., 1956), pp. 275–84.
[16] Tomasini, L. M., "Il Modello di Crescita di von Neumann ed Alcune Sue Generalizzazioni," L'Industria (1968), pp. 3–20.
[17] von Neumann, J., "Über ein ökonomisches Gleichungssystem und eine Verallgemeinereung des Brouwerschem Fixpunktsatzes," Ergebnisse eines mathematischen Kolloquiums, No. 8 (1937), pp. 73–83, translated as "A Model of General Economic Equilibrium" *Review of Economic Studies*, *12*, 1–9 (1945–6). Reprinted with corrections in *Precursors in Mathematical Economics: An Anthology*, Editors, W. J. Baumol and S. M. Goldfeld (London School of Economics, 1968), pp. 296–306.
[18] von Neumann, J. and O. Morgenstern, *Theory of Games and Economic Behavior* (Princeton University Press, Princeton, N.J., 1944), 3rd edition, 1953.

* * *

CONSUMPTION, THE RATE OF INTEREST AND THE RATE OF GROWTH IN THE VON NEUMANN MODEL*

Helmut Frisch

University of Vienna

1. INTRODUCTION

In this paper we study the role of consumption in the von Neumann growth model. In von Neumann's famous paper [13] consumption did not appear in an explicit form. The total output of period t is used up as input in $t+1$. One interpretation of this economy might be that the workers are slaves, so that their consumption in period t appears as input in $t+1$, or that the workers are consuming at [subsistence] level and the consumption of the entrepreneurs is zero. We call this concept, where consumption is considered as an input *the consumption-cost concept.* In this concept, the equilibrium condition of a growing economy of the von Neumann type is the equality of the rate of growth and the rate of interest.

A second approach would be to assume that consumption goods are produced for final consumption. They disappear and do not enter into the input matrix of the next period. We call this the *pure consumption concept.* For this case we show that in the state of balanced growth, the rate of interest must be greater than the rate of growth. The difference between these two rates shows the opportunity costs of consumption in terms of the rate of growth. We can show that the Morishima model [10] and the Malinvaud model [6] are variants of the "pure-consumption" approach. We also find that the Malinvaud version of the von Neumann model produces the same result as N. Kaldor's growth model, if one translates Kaldor's model into the von Neumann framework. We call this the Malinvaud-Kaldor version of the von Neumann model.

In a recent paper O. Morgenstern and G. Thompson [9] put the problem in a new light. They introduced the concept of income—instead of consumption—so the economic agents have the freedom of choice between consumption and saving. This produced the result that a growing economy obeys three regimes: growth rate $\gtreqless$ rate of interest, according to whether the consumption ratio is less than, equal to or larger than the savings ratio. Following the general approach of Morgenstern-Thompson, we add some details to this model, considering the types of consumption and savings functions, which have to be assumed; we also give a precise formulation of the equilibrium condition of the rate of growth and the rate of interest in this generalization.

1.1 Basic Features of the von Neumann Model

Let us first consider some basic features of the von Neumann model [1]. The technology is described by an input matrix A and an output-matrix B, both of the same dimension.

$$B=[b_{ij}]; \qquad A=[a_{ij}]; \qquad i=1, \ldots, m; \qquad j=1, \ldots, n.$$

*The research described in this paper was supported by the Ford Foundation, and the Econometric Research Program at Princeton University.

The input coefficient a_{ij} designates the amount of good j, which is used up in activity i, similarly b_{ij} denotes the output of good j from activity i, if the activity operates at unit intensity. We assume, that we have m activities and n goods in the economy.

There exists a $1 \times m$ row vector, called the intensity vector x, which is normalized, so the $\Sigma x_i = 1$. Each component x_i of this vector indicates the level at which activity i is operated. The $n \times 1$ column vector y represents a normalized price vector, where y_j stands for the price of a unit of commodity j. Both vectors are *semipositive*. The product

$$xA = \left(\sum_{i=1}^{m} a_{i1} x_i, \ . \ . \ . \ \sum_{i=1}^{m} a_{in} x_i \right) \tag{1}$$

is a row vector, that shows the total physical amounts of the goods $1, \ldots, n$, which are needed if the level of operation is x. The first component

$$\sum_{i=1}^{m} a_{i1} x_i$$

for example, is the amount of commodity 1 used up in the activities $1, \ldots, m$. A similar explanation holds for the product

$$xB = \left(\sum_{i=1}^{m} b_{i1} x_i, \ . \ . \ . \ . \ \sum_{i=1}^{m} b_{in} x_i \right), \tag{2}$$

which represents the total physical output of the n goods produced in the m activities of the economy.

Postmultiplying A by the column vector y, we get

$$Ay = \begin{pmatrix} \sum_{j=1}^{n} a_{1j} y_j \\ \vdots \\ \sum_{j=1}^{n} a_{mj} y_j \end{pmatrix} \tag{3}$$

Each component of this column vector denotes the input value of a process, if it is operated at the unit level. The product

$$By = \begin{pmatrix} \sum_{j=1}^{n} b_{1j} y_j \\ \vdots \\ \sum_{j=1}^{n} b_{mj} y_j \end{pmatrix} \tag{4}$$

defines the value of the output of each activity $1, \ldots, m$. Now we are able to consider the three important ratios in the von Neumann Economy.

Dividing the first component of the row vector (2) by the first component of the row vector in (1), the second component in (2) by the second component in (1), etc. until the n'th component, we get a set of ratios α_j as a function of x, the *expansion factor* of good j, $(j = 1, \ . \ . \ . \ n)$.

$$\alpha_j(x) = \frac{\sum_i b_{ij} x_i}{\sum_i a_{ij} x_i}. \tag{5}$$

If we divide in a similar way each component of the column vector (4) by the corresponding component of the column vector (3) we get a set of m ratios $\beta_i (i=1, \ldots m)$, measuring the profitability of the ith activity, as a function of y.

$$\beta_j(y)=\frac{\sum_i b_{ij}y_j}{\sum_j a_{ij}y_j}. \tag{6}$$

While the notion of the expansion factor presents no difficulties, it is important to keep in mind that β_i is nothing else than the ratio between the output value and the input value of process i. The total output value

$$\sum_i \sum_j b_{ij}x_i y_j;$$

divided by the total input value

$$\sum_i \sum_j a_i x_i y_j;$$

defines the monetary growth factor of the economy

$$\phi(x, y)=\frac{\sum_i \sum_j b_{ij}x_i y_j}{\sum_i \sum_j a_{ij}x_i y_j}. \tag{7}$$

1.2 Assumption About Technology

To facilitate our later discussion, we have to mention an important difference between the original von Neumann model [13] and the generalization of Kemeny, Morgenstern, and Thompson [4] with respect to the assumptions about technology.

Von Neumann's assumption was:

$$(^*)\ a_{ij}+b_{ij}>0 \qquad \text{for all } i \text{ and } j.$$

This assumption means that every process must either consume or produce a positive amount of every good. It is easy to see that this assumption is not too realistic, because it includes the case that we can produce something for nothing (if $a_{ij}=0$ and $b_{ij}>0$). This assumption was replaced by KMT by the intuitively more plausible concept:

(**) (*i*) every i of the input matrix A has at least one positive a_{ij}
(*ii*) every j of the output matrix B has at least one positive b_{ij}.

The economic implications of these assumptions are evident: (*i*) means that there is no activity i, without some positive input; (*ii*) every good in the economy can be produced by at least one activity.

1.3 The Axioms

Definition: A quadruple $\{x, y, \alpha, \beta\}$, where x is a normalized intensity vector, y a normalized price vector, α the expansions factor, and β the interest factor of the economy is called an N-equilibrium

if it satisfies Axioms I–IV and technology assumption (*), and a KMT-equilibrium if it satisfies Axioms I–V and technology assumption (**).

AXIOM I—Technological feasibility:

$$xB \geqq \alpha x A.$$

According to the first axiom the output vector, xB, in period t must be at least as large as the input vector, $\alpha x A$, in period $t+1$. If this holds, the economy can expand, with a constant (physical) growth factor α. Recalling the expressions (1) and (2) in 1.1 and (5) in 1.2, we see that the inequality holds for each component of the output and the input vector, so it follows that

$$\alpha \leqq \alpha_j(x)$$

where α_j is the expansion factor of the j-good. The total economy cannot grow faster than the slowest growing input factor j, that appears in the intensity vector x.

AXIOM II—Full Competition:

$$By \leqq \beta A y.$$

The second axiom makes the economy profitless. The monetary value of the output cannot be larger than the capitalized monetary input value in each activity. As stated by relation (6) we have for each of the m ratios β_i the inequality

$$\beta \geqq \beta_i(y), \qquad (i=1, \ldots, m).$$

There exists no activity in the economy with a larger profitability than β. The interest factor β is at least as large (in the weak sense) as the factor $\beta_i(y)$ of the most profitable activity i.

AXIOM III—No overproduction:

$$x(B-\alpha A)y=0.$$

The third axiom states that if in Axiom I the strict inequality holds, so that we are faced with overproduction, then the value of the overproduced goods should be zero. Overproduction means in this model that one or more goods in the economy grow faster than the total economy, so that for some j we have, the strict inequality

$$\alpha_j(x) > \alpha.$$

This implies, that the corresponding price $y_j=0$.

AXIOM IV—No Inefficient processes:

If on the other hand the inequality in Axiom II holds, there are some activities in the economy that are less profitable than the profitability of the total economy. These activities are inefficient and are used with zero intensity. More formally:

$$x(B-\beta A)y=0.$$

An activity i which is less profitable, than the interest factor:

$$\beta > \beta_i(y)$$

has a corresponding intensity $x_i=0$.

von Neumann proved in his famous paper that under Axioms I–IV and the technology assumption (*), an N-equilibrium always exists. If we use the technology assumption (**) we must impose Axiom V.

AXIOM V — Positive monetary output value:

$$xBy > 0.$$

The economic meaning of this axiom is trivial: something must be produced.

There exists an important difference between N-equilibria and KMT-equilibria. While the former are always unique, where $\alpha=\beta$, for the latter this is not generally true, in KMT-models we may have more than one equilibrium solution, characterized by the equation $\alpha=\beta$. It is possible to rank these solutions with respect to their values:

$$\alpha_1 > \alpha_2 \;.\;.\;. > \alpha_r,$$

where α_1 is the largest and α_r the smallest growth factor. For some of these α's Axiom V may not be satisfied. In these cases we have no "economic" solution. Thompson [12] proved that there always exists an economic solution for the largest and smallest α in the above row. In this case the economy is decomposable into self-sustaining subeconomies, and everyone of the equilibrium expansion factors α_k is the growth factor of at least one subeconomy. More precisely, we have the following existence theorem proved by KMT:

(A) If (**) holds, then there exists at least one and at most a finite number of α's for which the system I–V has economic solutions [4].

(B) (von Neumann). If (*) holds, then there is a unique α.

(C) If (*) and (**) hold, then there is an unique α moreover, for that α system I–V has economic solutions [4].

1.4 Game Theoretic Interpretation [1]

We consider the linear combination of the input and the output matrix $(B-\alpha A)$ as a payoff matrix of a matrix game, the probability vectors x and y are optimal mixed strategies. Write $M_\alpha = B-\alpha A$ for abbreviation; Axioms I–V now become [4]:

$$\begin{array}{rl} \mathrm{I}' & xM_\alpha \geqslant 0, \\ \mathrm{II}' & M_\beta y \leqslant 0, \\ \mathrm{III}' & xM_\alpha y = 0, \\ \mathrm{IV}' & xM_\beta y = 0, \\ \mathrm{V}' & xBy > 0. \end{array}$$

LEMMA 1: From this system we get the conclusion: if x, y, α, and β are solutions of I′–IV′ and V, then $\alpha=\beta$.

PROOF: From Axioms III′ and IV′, we have

$$xM_\alpha y = xM_\beta y$$

$$x(B-\alpha A)y = x(B-\beta A)y.$$

[1] We mention the game theoretic interpretation here, because this technique is used for solving the von Neumann model in part 4 of this paper.

Multiplying out, we get $(\beta-\alpha)xAy=0$. From condition (**) we know that the matrix A is not zero, so that $\alpha=\beta$ must hold. For equilibrium solutions $(\alpha=\beta)$ the system now shrinks to

$$\begin{array}{ll} \mathrm{I}' & xM_\alpha \geqq 0, \\ \mathrm{II}' & M_\alpha y \leqq 0, \\ \mathrm{V}' & xBy > 0. \end{array}$$

We omit Axiom III′ because it is contained in I′ and II′. Every triple $\{x \geqq 0,\ y \geqq 0,\ \alpha\}$ satisfying I′ and II′ must also satisfy III′.

PROOF: Multiplying (I′) by y gives $xM_\alpha y \geqq 0$, and (II′) by x gives $xM_\alpha y \leqq 0$; this implies $xM_\alpha y=0$, which is (III′).

In terms of game theory, we have the probability vectors, x and y, representing optimal mixed strategies, and the parameter α is a KMT= solution, if the value of this game

6.1 $$v(M_\alpha)=xM_\alpha y=0.$$

This suggests that the matrix game M_α is a "fair" game. For all other strategy vectors $\bar{x}$, $\bar{y}$ which are not optimal, we have the equilibrium of a saddle point:

6.2 $$\bar{x}M_\alpha y \leqq xM_\alpha y \leqq xM_\alpha \bar{y}.$$

PROOF: If x, y, and α is a KMT-solution, then we know that the value of the game $xM_\alpha y=0$. Multiplying I′ with $\bar{y}$ and II′ with $\bar{x}$ gives us $xM_\alpha\bar{y} \geqq 0$ and $\bar{x}M_\alpha y \leqq 0$.

COROLLARY: The normalized equilibrium vectors x (intensity vector) and y (price vector) are optimal strategies of the matrix game M_α, where α is a parameter. The saddlepoint (the value) of this two person zero sum game is zero.

1.5 A Two Good, Two Activity von Neumann Model

Consider, as an example, an economy producing two goods—consumption goods, K, and investment goods, P. The first row of the input and output matrix refers to the production of consumption-goods, the second to that of investment goods. The first column in the two matrices shows the production coefficients for consumption goods and the second the production coefficients for investment goods.

$$A=\begin{array}{c c} & \begin{array}{cc} K & P \end{array} \\ \begin{array}{c} K \\ P \end{array} & \begin{pmatrix} 1 & 3 \\ 2 & 1 \end{pmatrix} \end{array} \qquad B=\begin{array}{c c} & \begin{array}{cc} K & P \end{array} \\ \begin{array}{c} K \\ P \end{array} & \begin{pmatrix} 9 & 3 \\ 0 & 5 \end{pmatrix} \end{array}$$

At unit intensity we need one consumption good and three "machines" to produce an output of nine consumption goods and three (one-period-older) machines. The production good industry needs two consumption goods and one "machine" to produce a net output of four machines.

According to Axiom I, we have

(1) $$9x_1 \geqq \alpha(x_1+2x_2),$$

(2) $$3x_1+5x_2 \geqq \alpha(3x_1+x_2),$$

(3) $$x_1+x_2 = 1.$$

The objective is to maximize α, subject to contraints (1), (2), and (3). For the maximal α, we get

(4) $$\alpha_1 \leqq \frac{9x}{2-x},$$

(5) $$\alpha_2 \leqq \frac{5-2x}{1+2x}.$$

Solving (4) and (5), we get

$$\alpha = 2.29, \qquad x = (0.406,\ 0.594).$$

The maximal growth factor for this system is 2.29, the equilibrium intensity vector, shows that 0.406 percent of our total resources are used up for consumption good production and 0.594 percent for the production of investment goods.

With respect to Axiom II, we get

(6) $$9y_1 + 3y_2 \leqq \beta(y_1 + 3y_2),$$

(7) $$5y_2 \leqq \beta(2y_1 + y_2),$$

(8) $$y_1 + y_2 = 1.$$

(6) and (7) will be satisfied when β is large. The problem here is to find the minimal β, so that the three constraints (6), (7), and (8) are satisfied. The profitability of the consumption good activity is β_1, of the investment good activity β_2.

(9) $$\beta_1 \geqq \frac{6y+3}{3-2y}$$

(10) $$\beta_2 \geqq \frac{5-5y}{1+y}.$$

The solution of this problem is a price vector $y = (0.369,\ 0.631)$ which makes the profitability in both industries equal to

$$\beta = 2.3.$$

If the total price index is one, then the price of a consumption good is 0.369 and of a production good, 0.631. Figures 1 and 2 show the quantity and price relations in this economy. The shadowed area shows the region of the inequalities.

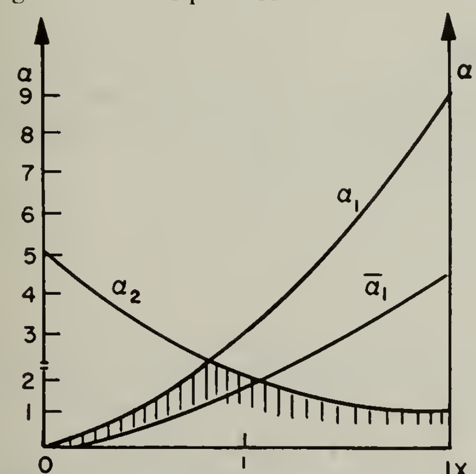


Fig. 1

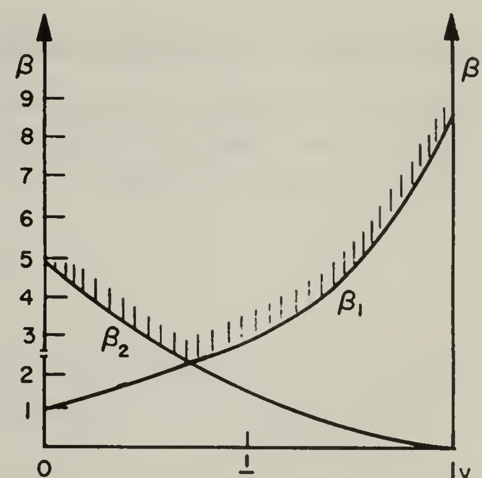


Fig. 2

According to Axioms III and IV, we have-

$$9x_1y_1+3x_1y_2+5x_2y_2-\alpha(x_1y_1+3x_1y_2+2x_2y_1+x_2y_2)=0$$
$$9x_1y_1+3x_1y_2+5x_2y_2-\beta(x_1y_1+3x_1y_2+2x_2y_1+x_2y_2)=0.$$

These two equations are satisfied for the equilibrium intensity $x=(0.406,\ 0.594)$, the equilibrium prices $y=(0.369,\ 0.631)$, and the equilibrium condition $\alpha=\beta=2.3$.
Axiom V says

$$xBy>0.$$

We have

$$(0.406,\ 0.594)\begin{pmatrix}9 & 3\\ 0 & 5\end{pmatrix}\begin{pmatrix}0.369\\ 0.631\end{pmatrix}=4.$$

This number is the total value which is produced in this economy.

Using the information which we obtained from the solution of this example we can examine this economy in more detail. Consider the two products:

$$Ay=\begin{pmatrix}2.262\\ 1.369\end{pmatrix}\qquad By=\begin{pmatrix}5.214\\ 3.155\end{pmatrix}.$$

The first column vector tells us that if we produce at unit intensity level, the value of the inputs in the consumption good industry is 2.262 and in the investment good industry 1.369; the value of the output of the consumption good industry is 5.214, the value of the output of the machine industry, 3.155. In both industries the profitability is 2.3, and both activities are used.

The two matrix products

$$xA=(1.594,\ 1.812);\qquad xB=(3.654,\ 4.188),$$

indicate that in equilibrium 1.6 units of consumption goods and 1.8 units of investment goods are used up as inputs. After one period the economy has 3.6 units of consumption and 4.2 units of investment goods. The equilibrium growth factor α_i for both commodities is 2.3.

2. CONSUMPTION IN THE VON NEUMANN MODEL

Introducing consumption or more accurately "extraconsumption" in the von Neumann model, we go back to Axiom I, which tells us that the output in period t must cover the input in period $t+1$. The output vector x_tB must be at least as large as the input vector $x_{t+1}A$, where

$$x_{t+1}=\alpha x_t$$
$$x_tB=\alpha x_tA.$$

If the society decides to consume, then for technological feasibility, the following inequality must hold:

$$x_tB\geqslant x_{t+1}A+x_{t+1}C,$$

where $C=(c_{ij})$ is the consumption matrix. The consumption coefficient c_{ij} designates the consumption of good j by the economic agents (workers and entrepreneurs) in activity i. The meaning of this inequality

is obvious: the output vector at time t must be at least as large as the input vector and the consumption vector in $t+1$. If we have n commodities, where $n-1$ are nonconsumption goods and the commodity n is the one consumption good, then we get, according to relation (5) in 1.2

$$\alpha_j = \frac{\sum_i b_{ij} x_i}{\sum_i a_{ij} x_i} \qquad j=1, \ldots, n-1,$$

and for the consumption good

$$\alpha_n = \frac{\sum_i b_{in} x_i}{\sum_i (a_{in} + c_{in}) x_i}.$$

In this case the growth factor of the total economy decreases in accordance with 1.3 Axiom I and coincides with the growth factor of the slowest growing commodity, the consumption good. *The effect of the introduction of consumption in the von Neumann model is a decline of the growth rate of the economy.*

We now return to the example in (1.5) and introduce a consumption matrix

$$C = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$$

so we have three matrices

$$A = \begin{pmatrix} 1 & 3 \\ 2 & 1 \end{pmatrix} \qquad C = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \qquad B = \begin{pmatrix} 9 & 3 \\ 0 & 5 \end{pmatrix},$$

with the following relations

(1) $$9x_1 \geqq \alpha_1[(x_1 + 2x_2) + (x_1 + x_2)]$$

(2) $$3x_1 + 5x_2 \geqq \alpha_2(3x_1 + x_2)$$

(3) $$x_1 + x_2 = 1.$$

The two growth factors are now

(4) $$\alpha_1 \leqq \frac{9x}{3-x}$$

and

(5) $$\alpha_2 \leqq \frac{5-2x}{1+2x}.$$

Consumption leads to a decline in the growth factor of the production of the consumption good. Maximizing α, with respect to (4) and (5) gives $\alpha = 1.92$ and the optimal intensity vector $x = (0.527, 0.472)$.

If we look at the products

$$xA + xC = (2.473, 2.055); \qquad xB = (4.746, 3.945),$$

we find that in equilibrium we use now 2.5 units of consumption goods and 2.5 units of investment goods as input in both activities, one period later we have 4.7 units of consumption—and 3.9 units of investment goods. We observe a shift in the allocation of resources to the consumption good industry compared to the example without consumption.

Figure 1 shows that the new function $\bar{\alpha}_1(x)$ shifted downward, while the graph of $\alpha_2(x)$ remains fixed. The new equilibrium growth factor is 1.92.

2.1 Concepts of Consumption

Considering Assumption II (full competition) we are faced with a critical problem: is consumption a factor of production, like the inputs in the input matrix A, and part of the costs of production, or is the economy a machine which produces consumption goods which do not enter into the costs of production? We call the first interpretation the "*consumption-cost*" concept; the second, the "*pure*" consumption concept.

The original von Neumann model [13], the discussion of "outside demand" in the KMT-model, [4] are specifications of the "consumption-cost" concept: consumption is treated like other costs of production. The Malinvaud model [6] and the Morishima version of the von Neumann model [10] are examples of the pure consumption concept: consumption goods are produced, but they are not part of the input vector.

Considering consumption as part of the production costs, the value of consumption has to be introduced in the input matrix of inequality II, and has to earn (as part of the unit costs) the equilibrium rate of interest $\beta-1$. Looking at the total system, we now have

$$\begin{array}{ll} \text{I}' & xB \geqq \alpha xA + \alpha xC, \\ \text{II}' & By \leqq \beta Ay + \beta Cy, \\ \text{III}' & x[B-\alpha(A+C)]y = 0, \\ \text{IV}' & x[B-\beta(A+C)]y = 0, \\ \text{V}. & xBy > 0. \end{array}$$

A quadruple $\{x, y, \alpha, \beta\}$ which satisfies these five axioms is a generalized von Neumann equilibrium. It is obvious that for equilibrium we have $\alpha=\beta$. The only difference from the model without consumption is, that the growth factor α and the interest factor β are both lower than in the case without consumption. There will be different equilibrium values for x and y.

Introducing the pure consumption concept into the model, we now prove that in equilibrium the rate of interest must be larger than the rate of growth. More precisely: *The ratio of the interest-factor β and the growth factor α minus one is equal to the ratio of the value of total consumption to the value of the total input (the consumption input ratio: xCy/xAy*

In deriving this Theorem we use the same axioms and technique as before. The economy is now described by an input matrix A, an output matrix B and a consumption matrix C. Consider the following system:

Axiom I″: $x(B-\alpha A) \geqq \alpha xC$.

The difference between the output vector $x_t B$ and the input vector $x_{t+1}A$ must be at least sufficient to satisfy the demand for consumption goods $x_{t+1}C$.

Axiom II″: $(B-\beta A)y \leqq 0$.

According to the pure consumption concept, the introduction of consumption has no influence on the profitability of production.

Axiom III″: $x[(B-\alpha(A+C)]y=0$.

If from commodity j in period t, is produced more than that which is consumed in period $t+1$, or that which goes into the input vector in $t+1$, then the price of this commodity y_j is zero. Overproduction is the difference between the amount of commodity j available and that which is used as input or for consumption. The value of overproduction is zero.

Axiom IV: $x(B-\beta A)y=0$.

Axiom IV stays unaltered. This is the assumption, that consumption is not a part of the costs of production. If the strict inequality $\beta>\beta_i(y)$ holds, the process i has a zero intensity.

We have also this last axiom:

Axiom V: $xBy>0$.

PROPOSITION: If x, y, α, and β are solutions of the system $\mathrm{I}'-\mathrm{V}$, then

$$\frac{\beta}{\alpha}-1=\frac{xCy}{xAy}.$$

PROOF: From Equations III′ and IV we get

$$\alpha xAy+\alpha xCy=\beta xAy,$$

this gives:

$$(\beta-\alpha)xAy=\alpha xCy.$$

From assumption (**) and Assumption I″ we know that $xAy>0$ and $xCy>0$, so that the relation $\beta>\alpha$ must hold. Dividing the equation by the scalar xAy, gives the above result. The equation

$$\frac{\beta}{\alpha}-1=\frac{xCy}{xAy}$$

shows that if we introduce consumption into the von Neumann economy in the sense of the pure consumption concept, then the total amount of interest-cost of the society $(\beta-1)xAy$ is larger than the increase in the output value $(\alpha-1)xAy$ of the economy. The ratio of the interest rate and the growth rate gives a natural measure of the opportunity cost to society. The higher consumption, the larger

$$\frac{\beta}{\alpha}-1$$

and the larger the opportunity cost of consumption in terms of the rate of growth of the economy. Consider two cases to see the economic logic of this situation: if xCy is such that

$$\alpha=\frac{xBy}{xAy+xCy}=1,$$

so that we have a stationary economy, then the formula gives us $(\beta-1)xAy=xCy$. $\beta-1$ represents the rate of interest and $(\beta-1)xAy$ is the amount of "profit" which the economy earns. The total profit of the economy is consumed, no growth can take place.

If on the other hand, $\alpha=\beta$, consumption must be zero. The economy has no opportunity costs in terms of the rate of growth.

2.2 Example

For the system I′ − V, we get the following relations in our 2 x 2 economy.

I′: (I′.1) $9x_1 \geqq \alpha[(x_1+2x_2)+(x_1+x_2)]$,
(I′.2) $3x_1+5x_2 \geqq \alpha(3x_1+x_2)$,
(I′.3) $x_1+x_2=1$.

The growth factor α is maximized, if $\alpha=1.92$ and $x=(0.5274,\ 0.4726)$.

II′: (II′.1) $9y_1+3y_2 \leqq \beta(y_1+3y_2)$,
(II′.2) $5y_2 \leqq \beta(2y_1+y_2)$,
(II′.3) $y_1+y_2=1$.

The interest factor is minimized if $\beta=2.3$ and the price vector $y=(0.369,\ 0.631)$.

III: $9x_1y_1+3x_1y_2+5x_2y_2-\alpha(2x_1y_1+3x_1y_2+3x_2y_1+x_2y_2)=0$.

IV: $9x_1y_1+3x_1y_2+5x_2y_2-\beta(x_1y_1+3x_1y_2+2x_2y_1+x_2y_2)=0$.

Solving these two equations, we get

$$\beta=\alpha+\alpha\frac{(x_1y_1+x_2y_1)}{x_1y_1+3x_1y_2+2x_2y_1+x_2y_2},$$

or by using I′.3 and II′.3

$$\beta=\alpha+\frac{\alpha y}{1+2x+y-3xy}.$$

Checking the result by using the equilibrium values $x=0.527$, $y=0.369$, and $\alpha=1.92$, we find that $\beta=2.3$. The total value of the output xBy is

$$(0.5274,\ 0.4726)\begin{pmatrix}9 & 3\\ 0 & 5\end{pmatrix}\begin{pmatrix}0.369\\ 0.631\end{pmatrix}'=4.2.$$

3. THE MORISHIMA AND MALINVAUD VERSION

We do not intend to review the Morishima and Malinvaud model completely. We consider only that part of their models connected with the problem of consumption and the rate of interest. It may be convenient to discuss first some new concepts which are introduced by Morishima. We then construct, using these concepts, a von Neumann model along the lines of the previous section and show that the results of Morishima and Malinvaud are special variants of our earlier theorem.

3.1 The von Neumann-Morishima Model

The main contribution of Morishima [10] is the introduction of labor in an explicit form into the model. Following him, we introduce a vector of labor coefficients and the explicit price of a unit of labor, the wage rate. Morishima assumes that the economy has a homogeneous labor force, so that the real wage rate,

$$\omega_t = \frac{\omega_n}{\sum_i y_i},$$

is equal for all workers in every industry. If there are m activities in the economy we have a column vector of m labor coefficients:

$$L = \begin{pmatrix} \ell_1 \\ \ell_2 \\ \cdot \\ \cdot \\ \cdot \\ \ell_m \end{pmatrix}.$$

The scalar product

$$x \cdot L = \sum_{i=1}^{m} x_i \ell_i$$

is the total labor requirement. Multiplying this figure by the outside given real wage rate $\omega_t \cdot x \cdot L$, we get the total labor cost or labor income of the economy. Morishima's model is a two-class model. The income is distributed between the class of workers who receive a wage payment and the class of investers who earn a pure interest income. Both have different consumption functions, but each is homogeneous of degree zero with respect to the price vector y and income.

The n-dimensional row vector

$$d_t = (d_1, \ldots, d_n)$$

designates, the demand of investors for the goods $1 \ldots, n$ in period t. The ratios $d_1 : d_2 : \ldots d_n$ are independent of the level of income u_t of the investors. The market demand function for good j_j is given by:

$$d_j = f_j(y) \cdot u_t \; j = 1, \ldots, n$$

if $u_t \geqslant 0$, and $d_j = 0$ if $u_t \leqslant 0$. $f(y)$ is an n-dimensional row vector, such that:

$$\sum_j f_j(y) \cdot y_j = c,$$

where c is a constant and $0 < c < 1$.

Summing we get the total demand of the investors

$$\sum_j c_j y_j = u_t \sum_j f_j(y) \cdot y_j$$

or

$$d_t \cdot y_t = c \cdot u_t.$$

The income of the investors is available at the end of a production period, so we have, with respect to the time index,

$$u_{t+1} = (\beta - 1) x_t A y_t$$

where

$$u_{t+1} = \alpha u_t.$$

Let an n-dimensional row vector e_t

$$e_t = (e_1, \ . \ . \ . \ , e_n)$$

represent workers consumption in period t, then we have

$$W_{(t)} = e_t \cdot y_t.$$

The total labor income $\omega_t \cdot x_t \cdot L = W_{(t)}$ equals the total sum of their expenditure for consumption goods

$$\sum_j e_j y_j,$$

the saving ratio of the workers is zero. The workers market demand for good j, is given by the equation:

$$e_j = g_j(y) \cdot W_{(t)} j = 1, \ . \ . \ . \ , n.$$

There exists also an n-dimensional row vector $g(y)$ such that

$$\sum_j g_j(y) \cdot y_j = 1.$$

This is another way of saying that the consumption-income ratio of the workers is one; the sum of the proportional nominal spending for the goods $j = 1, \ . \ . \ . \ , n$ must be one.

$$\sum_j g_j(y) y_j = \frac{1}{W} \sum_j e_j y_j = 1.$$

The total demand of the wage earners is then

$$\sum_j e_j y_j = W_t \cdot \sum_j g_j(y) y_j$$

or

$$e_t \cdot y_t = W_t.$$

Morishima distinguishes two versions of this model: the difference is praenumerando or postnumerando payment of the wage bill. In the first case the workers receive their income at the beginning of the production period, in the second case at the end. In the latter case, the capital that the entrepreneurs invest for the labor force will not earn an interest payment. Morishima designates the first version, the Marx-von Neumann version, the second the Walras-von Neumann version. We consider only the Walras-von Neumann version. Here we have the relation

$$W_{t+1} = w_t \cdot x_t L.$$

The laborers consume in $t+1$, what they earn at the end of period t:

$$\alpha e_t y_t = w_t x_t L.$$

It may be helpful to remember Walras' definition of the unit cost of an industry i. If a_{ij} is the production coefficient of good j in activity i and τ_{ij} the length of life of capital good j in process i, the Walras unit cost is:

$$\sum_{j=1}^{n} (a_{ij}/\tau_{ij})\cdot y_j + r_{(t)} \cdot \sum_{j=1}^{n} a_{ij}y_j + \omega\cdot L.$$

In a von Neumann economy, goods at different stages of wear and tear are different goods, so all τ_{ij} are equal to unity. Thus we have for the process i

$$\beta_t \sum_{j=1}^{n} a_{ij}y_j + w_t\cdot L,$$

where $\beta_t = 1 + r_\tau$ the interest factor. The labor costs are paid at the end of the period, so they are not multiplied by the interest factor β_t.

The Morishima-Walras version of the von Neumann model is now described by the following relations:

$$xB \geqq \alpha xA + \alpha e + \alpha d \tag{1}$$

$$By \leqq \beta Ay + wL \tag{2}$$

$$xBy = \alpha xAy + \alpha ey + \alpha dy \tag{3}$$

$$xBy = \beta xAy + wxL \tag{4}$$

$$x \geqq 0,\ y \geqq 0,\ xBy > 0. \tag{5}$$

The equilibrium relation between the rate of interest and the rate of growth is, in the MW-version of the von Neumann model:

$$\begin{cases} \beta = \alpha + c(\beta - 1) & \text{if } \beta \geqq 1 \\ \beta = \alpha & \text{if } \beta \leqq 1. \end{cases} \tag{6}$$

PROOF: From (5) and (4), $xBy > 0$.

From relation (3) and (4), we find

$$(\beta - \alpha)xAy = \alpha dy,$$

because $\alpha ey = wxL$; we know that $\alpha dy = c(\beta - 1)xAy$, so that relation (6) follows.

Another direct way to get Morishima's result would be to use the formula from 2.1

$$\frac{\beta}{\alpha} - 1 = \frac{xCy}{xAy}$$

and the definitions of the demand in Morishima's model.

We know that the consumption vector, which appears in Axiom I and not in Axiom II makes a positive difference between β and α. The consumption of the workers appears in both relations as demand for consumption goods (in Eq. I) and as labor costs (wage bill) in Eq. II. The reason for the difference is the "pure" consumption of the investors

$$xCy = dy.$$

The n-dimensional demand vector xC corresponds to the n-dimensional demand vector d. In period t the investors spend the portion c from their income earned in period $t-1$.

$$\alpha xCy = \alpha dy = c \cdot (\beta-1)xAy.$$

Using this relation, we get

$$\frac{\beta}{\alpha} - 1 = c\frac{(\beta-1)}{\alpha},$$

where $(\beta-1)$ is the rate of interest (or rate of profit) and $(\beta-1)\cdot c$ is the opportunity cost of the extra consumption of the entrepreneurs. Rearranging gives a well-known result in growth theory:

$$\alpha - 1 = s(\beta-1).$$

The rate of growth $\alpha-1$ is proportional to the rate of interest, where s is a constant of proportionality. For $s=1$, the opportunity-costs of consumption are zero and the economy expands with a maximal growth rate $\alpha-1=\beta-1$.

3.2 The von Neumann-Malinvaud Model

Malinvaud [6] considers the uniform and maximal expansion of an economy under the following conditions. There exists a von Neumann economy with an input matrix $A=(a_{ij})$ and an output matrix $B=(b_{ij})$ and

$$A \geqq 0 \qquad B \geqq 0.$$

For period t there exists an n-dimensional row vector b_t with the relation

$$b_t = a_t + d_t,$$

where b_t is the output vector, a_t the input vector, and d_t the vector of the final demand or consumption. There exists also, an m-dimensional row vector x, that denotes the intensity of the process.

A uniform "expansion program" is characterized by the growth factor $\alpha \geqslant 0$, three n-dimensional row vectors a, b, d, and an m-dimensional row vector x; (α, a, b, d, x) with the equations:

$$a_t = \alpha^t a \qquad d_t = \alpha^t d$$

$$b_t = \alpha^t b \qquad x_t = \alpha^t x.$$

Uniform expansion (for every t) requires the following set of relations:

$$\begin{aligned} b &= a + d \\ a &\geqq xA \\ \alpha b &\leqq xB \\ x &\geqq 0,\ a \geqq 0,\ b \geqq 0,\ \alpha \geqq 0. \end{aligned}$$

This system implies the inequalities

$$\alpha xA + \alpha d \leqq \alpha a + \alpha d = \alpha b \leqq xB.$$

Rearranging this gives

$$x(B-\alpha A) \geqq \alpha d \tag{1}$$

$$x \geqq 0, \qquad \alpha > 0.$$

We know this relation from the previous sections. The economic meaning of the inequality is that the consumption vector d in period $t+1$ cannot be larger than the vector difference between output in period t and inputs in period $t+1$. We also observe that relation (1) becomes an equation if

$$(a-xA)=0 \qquad \text{and} \qquad (xB-\alpha b)=0,$$

which means that there exists no input which is left over, and no production that exceeds the output. Two expansions with the same triple (α, x, d) are not distinct.

DEFINITION: A uniform expansion (α, x, d) is h-efficient if there exists no h-program, that satisfies (1) and

$$\begin{aligned} \bar{d}_t &\geqq \alpha^t d \\ a_0 &\leqq a \qquad t=1, 2, \ldots, h-1, \\ \bar{b}_h &\geqq \alpha^h b \end{aligned}$$

where in at least one of these $h+1$ inequalities, strong inequality holds. An expansion program is h-efficient if it is not dominated by another h-program.

For the consumption vector d, Malinvaud assumes that

(2) $$d=xD,$$

where D is an $m\times n$ matrix and corresponds to the consumption matrix of the previous section.

PROPOSITION (1): The uniform expansion (α, x, d) is maximal, if there exists an m-dimensional column vector $y>0$, with

$$(B-\alpha A)y \leqq \alpha Dy.$$

PROOF: If $(\bar{\alpha}, \bar{x}, \bar{d})$ is a uniform expansion, then we have from (1) and (2)

(3) $$\bar{x}[B-\bar{\alpha}(A+D)] \geqq 0.$$

Because $\bar{y}>0$ and $\bar{\alpha}>0$, we can write

(4) $$\frac{1}{\bar{\alpha}}\bar{x}By-\bar{x}(A+D)y \leqq 0.$$

From Proposition 1, we have the relation

(5) $$[B-\alpha(A+D)]y \leqq 0;$$

because of $\bar{x} \geqslant 0$ and $\alpha>0$, we have

(6) $$\frac{1}{a}\bar{x}By-\bar{x}(A+D)y \leqq 0.$$

Subtracting (4) and (6), we get

$$\left(\frac{1}{\bar{\alpha}}-\frac{1}{\alpha}\right)\bar{x}By \geqq 0;$$

we have $\bar{x} \geqq 0$, $y \geqq 0$, $B \geqq 0$; $\Rightarrow \bar{x}By>0$

$$\left[\left(\frac{1}{\bar{\alpha}}\right)-\left(\frac{1}{\alpha}\right)\right] \geqslant 0$$

it follows that $\bar{\alpha} \leqq \alpha$, and α is maximal [7]

PROPOSITION (2): If a uniform expansion (α, x, d), with $v(B) > 0$ is h-efficient $(t=1, \ldots, h)$, then there exists a column vector $y \geqslant 0$ and a number $\beta > 0$ with

$$x(B-\alpha A)=\alpha d \tag{7}$$

$$(B-\beta A)y \leqq 0 \tag{8}$$

$$x(B-\beta A)y=0. \tag{9}$$

PROOF (for (7)):

The n-dimensional row vector $\delta = x(B-\alpha A)-\alpha d$ is semi-positive, because of (1). Rewriting we get

$$x(B-\alpha A)=\alpha\left(d+\frac{\delta}{\alpha}\right).$$

The triple $\left(\alpha, \bar{x}, d+\frac{\delta}{\alpha}\right)$ is also a uniform expansion. Malinvaud shows that this program is not h-efficient. There must be a program where

$$\bar{a}_t=\alpha^t xA \qquad \bar{b}_t=\alpha^{t+1}\cdot xB$$

$$\bar{d}_t=\alpha^t\left(d+\frac{\delta}{\alpha}\right) \qquad \bar{x}_t=\alpha^t\cdot x.$$

From the fact that (α, x, d) are h-efficient and the definition of h-efficiency, we obtain

$$\bar{d}_t \leqq \alpha^t d$$

or

$$\alpha^t\left(d+\frac{\delta}{\alpha}\right) \leqq \alpha^t d.$$

This relation is only fulfilled for $\delta=0$; and we get (7). Malinvaud considers

$$q=Ay \qquad \text{and} \qquad h=\frac{d}{xAy}. \tag{10}$$

The m-dimensional row vector q shows the monetary value of the inputs of the m activities, the m-dimensional column vector h, the relation between the demand for the n goods and the total input value.

If (α, x, d) are maximal, then we have with respect to Proposition 1 and (10)

$$x(B-\alpha A) \geqq \alpha d, \tag{11}$$

$$(B-\alpha A)y \leqq \alpha q\cdot hy, \tag{12}$$

$$x(B-\alpha A)y=\alpha dy. \tag{13}$$

We have to remark that the product of the $(1\times m)$ vector q and the $(m\times 1)$ vector h; $q\cdot h=D$.

We obtain the n-dimensional consumption (demand vector)

$$d=(d_1, \ldots, d_n) \quad \text{from} \quad x\cdot q\cdot h=xD=d.$$

It is easy to see, that

$$dy=\sum_{j=1}^{n} d_j y_j$$

is the total value of expenditures for consumption goods.

PROPOSITION (3):

For a program (α, x, d), that is h-efficient and maximal, we have the relation

$$\beta-\alpha=\alpha\frac{dy}{xAy}=\alpha\cdot h\,y, \tag{14}$$

$$\beta=\alpha(1+h\,y). \tag{15}$$

PROOF: The relation (7) implies relation (11) and (13).
From Eq. (9) and (13), we obtain

$$\beta xAy=\alpha xAy+\alpha dy.$$

From this relation (14) and (15) follow, taking (10) into account.

From (8) subtracting $\alpha\,A\,y$ on both sides, one obtains

$$By-\alpha Ay\leqq\beta Ay-\alpha Ay\leqq(\beta-\alpha)\,Ay.$$

Using (14) we get

$$(B-\alpha A)\;y\leqq\alpha\frac{dy}{xAy}\cdot Ay.$$

This leads, by use of (10) and (11), to (12) $(B-\alpha A)\;y\leqq\alpha\,q\,h\,y$. From proposition 1, we know that (α, x, d) is maximal.

3.3 The von Neumann-Kaldor Model

In this section we consider the disaggregation which Malinvaud introduced into the von Neumann model, also his introduction of a consumption-function of the entrepreneurs and of the "Facteurs primaires" (labor). Malinvaud's result, the equilibrium relation between the rate of growth, the rate of interest, the saving ratios and the distribution of income between entrepreneurs and the primary factors of production is, as we will show, identical with the well known result of N. Kaldor's growth model [3, 11].

Malinvaud distinguishes the following set of n-dimensional row vectors:

$c=(c_j)$	consumption,
$d=(d_j)$	final demand,
$f=(f_j)$	primary factors of production,
$k=(k_j)$	real capital,
$a=(a_j)$	total input,
$b=(b_j)$	total output.

Every row vector is the product of the $(1\times n)$ dimensional intensity vector, x, times the respective matrix. For example $k=x\cdot K$, where $K=(k_{ij})$ and the component, k_{ij}, denotes the amount of capital good, j, used in the activity, i.

For these six vectors, we have the following three equations:

(1) $c=d+f$ (consumption = final demand + input of primary factors),

(2) $k=a-f$ (real capital = total input − input of primary factors),

(3) $k=b-c$ (real capital = total output − consumption).

From (1), (2), and (3), we get

$$b=a+d=c+k. \tag{4}$$

The interest income of the entrepreneurs (investors) is now

$$u=(\beta-1)\,ky,$$

and the income of the primary factors of production is

$$w=\beta fy.$$

Malinvaud assumes that both types of economic agents have a positive but different propensity to save:

$$0<s_u\leq 1;$$

$$0<s_w\leq 1; \text{ and } s_u\neq s_w.$$

The monetary value of the total consumption of the society in period $t+1$ is now

$$\alpha cy=c_u(\beta-1)ky+c_w\beta fy. \tag{5}$$

The value of savings is

$$u+w-\alpha\, c\, y=(1-c_u)\,(\beta-1)\,ky+(1-c_w)\,\beta fy \tag{6}$$

or

$$(\beta-1)\,ky+\beta fy-\alpha\, cy=s_u(\beta-1)\,ky+s_w\beta fy. \tag{7}$$

By rearranging (7) using (5), we get

$$(\alpha-1)\,ky+\beta\, ay-\alpha\, by=s_u(\beta-1)\,ky+s_w\beta fy. \tag{8}$$

For an efficient growth program, the relations

$$a=xA,\ \alpha b=xB, \text{ and } x\,By=\beta xAy$$

must be satisfied, so we obtain

$$\frac{\alpha-1}{\beta-1}=s_u+s_w\frac{w}{u}. \tag{9}$$

Malinvaud's result [8].

It is not difficult to see that Malinvaud's result is identical with N. Kaldor's formula for the growth rate in an aggregated neo-Keynesian-growth model. In this model the rate of capital accumulation is connected with the distribution of incomes between the entrepreneurs and the laborors. We have to rearrange Malinvaud's formula in this way:

The net social product in a von Neumann economy consists of the aggregated income of both types of economic agents:

$$(\beta-1)\,ky+\beta\, fy=X, \tag{10}$$

where X is a single real number expressing the (equilibrium) value of the net production of the economy. Using identity (10), we obtain from (9)

$$\alpha - 1 = (s_u - s_w)\ (\beta - 1) + s_w \frac{X}{ky}, \tag{11}$$

which is exactly N. Kaldor's growth formula, if we translate and write for the rate of growth of the capital stock $dK/dt \cdot 1/K$, for the rate of profit P/K, and for the capital productivity X/K, giving

$$\frac{dK}{dt} \cdot \frac{1}{K} = (s_u - s_w)\ \frac{P}{K} + s_w \frac{X}{K}. \tag{12}$$

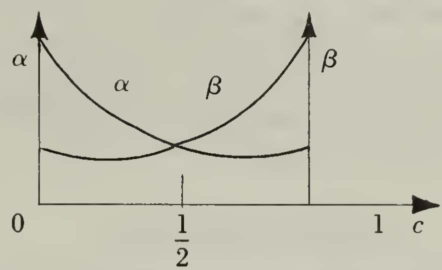
4. THE MORGENSTERN-THOMPSON VERSION OF THE VON NEUMANN MODEL

An important step in the consideration of the relation between consumption, savings, and the rate of interest is made in the Morgenstern-Thompson version of the von Neumann model [9].

The difference between this and the other models is the fact that they introduce the concept of income, a consumption, and a saving function into the economy. In every period the economy produces an income and this is partly consumed and partly invested. For such an economy they prove that we have the following relations between the savings ratios, the consumption ratio c, α, and β:

$$\begin{aligned} \beta &> \alpha \qquad \text{if } c > s \\ \alpha &> \beta \qquad \text{if } s > c \\ \alpha &= \beta \qquad \text{if } s = c. \end{aligned}$$

By drawing a graph of the relations, we get



where
$$\alpha = \frac{xBy}{x(A + cW)y}, \qquad \beta = \frac{xBy}{x(A + sW)y},$$

and the matrix $W = (w_{ij})$, where w_{ij} is the amount of good, j, given to the worker in the activity, i, if the industry is run at unit intensity level.

The economic meaning of this result is: If the consumption ratio of the society is larger than the saving ratio, the rate of interest exceeds the rate of growth; if the saving ratio exceeds the consumption ratio then the inverse relation holds, the rate of growth is larger than the rate of interest. Both rates are equal if the savings-ratio is equal the consumption ratio.

We reconsider this model, giving a more general interpretation to the matrix W and also showing the precise equilibrium relationship between β and α. Also we will go more explicitly through the details of this model.

We have to introduce the concept of income into the von Neumann model. The basic concept is the income matrix $E = (e_{ij})$ $i = 1, \ldots, m, j = 1, \ldots, n$, where the component e_{ij} shows the amount of good, j, that is obtained as income by the economic agents (workers and entrepreneurs) in activity, i, if the level of operation is one. These agents make decisions as to how much of this amount they

should save (= invest) and how much they should consume. The product, $I=Ey$, where y is an $(m\times 1)$ dimensional column vector, shows the value of income in every one of the m industries. For example in industry 1, the value of income (if the intensity of activity is one) is

$$I_1=\sum_{j=1}^{n} e_{1j}y_j.$$

The product $e=xE$ is a $(1\times n)$ dimensional row vector, where the component

$$e_j=\sum_i x_i\, e_{ij}$$

denotes the (physical) amount of good. j, that is obtained as income by the economic agents in the m-activities. The total value of income is expressed by

$$xI \text{ or } ey, =\sum_i\sum_j x_i\, e_{ij}y_j.$$

The market demand functions of the economy are homogeneous of degree zero, with respect to the income, ey, and the price vector, y, so we avoid money illusion of the economic agents. We also assume that the elasticity of demand for good, j, with respect to income is one; that implies that the structure of demand; d_1: d_2 . . . : d_n is not a function of the level of income.

The demand function for good, j, is described by

(1) $$d_j=d_j(y)\cdot ey,$$

where $d_j(y)$ is the proportion of the demand for good j, with respect to the total income ey. The demand for the j goods in each period is represented by the vector $d=(d_j)$. Multiplying the $(1\times n)$ dimensional vector $d=(d_j)$ and $d(y)=(d_i(y))$ with the price vector y, we get

(2) $$\sum_{j=1}^{n} d_jy_j=\sum_{j=1}^{n} d_j(y)y_j\cdot ey.$$

We assume that the average propensity to consume $0<c\leqq 1$ is constant, so we stipulate that

(3) $$\Sigma d_j(y)y_j=c.$$

From (2), we see that the total demand value for the n goods is equal to that part of income that is spent for consumption.

(4) $$\sum_j d_jy_j=c\cdot ey.$$

From the first axiom, expressing the *technological feasibility of consumption*, we get

(5) $$x(B-\alpha A)\geqq \alpha d.$$

According to (5) we have, for good j the relation

$$\sum_{i=1}^{m} b_{ij}x_i-\alpha\sum_{i=1}^{m} a_{ij}x_i\geqq \alpha d_j,$$

where
$$d_j = c \cdot e_j = c \sum_i e_{ij} x_i.$$

The m-dimensional column vector $I = Ey$, gives us information about the value of income that is created in the m-processes, if the intensity level of production is one.

$$I_i = \left(\sum_j e_{ij} y_j\right) \qquad (i = 1, \ldots, m).$$

From this income the constant proportion s, is saved in every period. In every industry, i, the amount

$$s \sum_j e_{ij} y_j$$

is invested* which will earn an amount of interest

$$(\beta - 1)s \cdot \sum_j e_{ij} y_j.$$

The unit costs of production for industry i are now

$$\beta \sum_j a_{ij} y_j + \beta s \sum_j e_{ij} y_j.$$

Axiom II stipulates "full competition" (no profit) so we can write the m inequalities in matrix form:

$$(B - \beta A) y \leqq \beta s I.$$

This expresses the fact that the value of output is less than or equal to the total capitalized value of inputs *and* savings. Axiom III states that there is no overproduction. If from a good, j, in period, t, is produced more than is used up as input or consumption in $t+1$, so that

$$\sum_{i=1}^{m} b_{ij} x_i - \alpha \left(\sum_{i=1}^{m} a_{ij} x_i + d_j\right) > 0,$$

then y_j is zero. So we have

$$x[B - \alpha(A + cE)]y = 0.$$

Axiom IV states that there are no inefficient processes, which means that if the unit value of output of activity, i,

$$\sum_{j=1}^{n} b_{ij} y_j - \beta \left(\sum_j a_{ij} y_j + s \Sigma e_{ij} y_j\right) < 0$$

does not cover the capitalized value of inputs and savings then the respective activity level is zero:

$$x[B - \beta(A + sE)]y = 0.$$

Axiom V says that the value of the total output must be greater than zero,

$$xBy > 0.$$

The technological assumptions:

(**) For the input matrix, we have:
for every i exists one j with $a_{ij} > 0$.

*Savings equals investment in the classical equilibrium sense, not in Keynes' tautological sense.

For the output matrix:
for every j exists one i with $b_{ij} > 0$.

$$a_{ij} + b_{ij} + e_{ij} > 0. \tag{*}$$

The assumption (**) is the KMT-assumption and means there exists no row of zeros in the input matrix, and no column of zeros in the output matrix. The consequence of this assumption is that the economy has now more than one equilibrium path, each corresponding to a sub-economy of a set of activities and a set of commodities. To avoid this breaking up of the economy into sub-economies, M–T introduced (*), which is very similar to von Neumann's original assumption

$$a_{ij} + b_{ij} > 0.$$

LEMMA: If $x \geqq 0$, $y \geqq 0$, α and β are solutions (a $M-T$ equilibrium) of the system I–V, then

$$\frac{\beta}{\alpha} = \frac{1+cp}{1+sp}, \tag{6}$$

where
$$p = \frac{ey}{xAy},$$

the ratio between the value of income and the value of capital, the "capital productivity" in the von Neumann model.

PROOF: From III, we get $xBy = \alpha(xAy + cey)$, from IV, $xBy = \beta(xAy + sxI)$. From V, we know that $xBy > 0$, so all terms in both brackets cannot be zero. We get

$$\beta(xAy + sxI) = \alpha(xAy + cey).$$

By dividing both sides by xAy, and remembering that the monetary income

$$\sum_i \sum_j x_i e_{ij} y_j = xI = ey.$$

one gets the above relation. By inspecting the equilibrium relation, we find that the ratio β/α depends for a given value of the "capital coefficient" on the ratio of the consumption and the savings coefficient. There are three regimes of the economy

$$c > s => \beta > \alpha \tag{1}$$

$$c < s => \beta < \alpha \tag{2}$$

$$c = s => \beta = \alpha. \tag{3}$$

Regime (1) describes an economy with a positive time preference, that is measured by the difference $\beta - \alpha$. The economic agents only save, if they receive a positive ratio above the rate of growth. The consumption ratio exceeds the saving ratio. Regime (2) refers to a situation in which the time preference of society is negative. The economic agents prefer future consumption to present consumption and will save even if the rate of interest is less than the rate of growth. The goal is to transfer present income

to the future. In this situation the saving ratio is larger than the consumption-ratio. Regime (3) is characterized by a rate of time preference of zero, so that the rate of interest equals the rate of growth.

Morgenstern and Thompson mentioned that we can use III and IV in defining two separate matrix games:

$$M(\alpha)=B-\alpha\,(A+c\,E)$$
$$M(\beta)=B-\beta\,(A+s\,E).$$

Each of these games can be considered, using the assumptions of a von Neumann economy, as an expanding economy. Each of these economies can be characterized by a triple (x, y, α), where $\alpha=\beta$ in each case; let us call M_α the quantity and M_β the price game.

Using a formulation of von Neumann [2] and [13], we are faced with the following problems:

(1) A growth factor α and an intensity system $x \geqq 0$ are equilibrium solutions, if α and x are solutions of the problem: Find the maximal α, under the constraints:

$$xB \geqq \alpha x(A+cE) \qquad (5.1)$$
$$x \geqq 0.$$

(2) The interest factor β and the price vector $y \geqq 0$, are equilibrium solutions, if α and y are solutions of the problem: Find the maximal β under the constraints:

$$By \leqq \beta(A+sE)y \qquad (5.2)$$
$$y \geqq 0.$$

PROPOSITION 1: If there is a solution α' to the quantity game M_α, then α' is the maximal α of (5.1). If there is a solution β' to the price game M_β, then β' is the minimal β of (5.2).

PROOF: In section 1.6, a KMT-solution was defined by $(x \geqq 0,\ y \geqq 0,\ \alpha')$ with $xM_\alpha y=0$. Assume that there exists an $\alpha > \alpha'$. Since the matrices A and E are nonnegative, $c \geqq 0$ and α' unique, the assumption $\alpha > \alpha'$ implies $xM_\alpha \leqq 0$, which contradicts to assumption (5.1), so $\alpha \leqq \alpha'$. Assume there exists $\beta < \beta'$. This would imply, because of $v(M(\beta'))=0$ $M_\beta y \geqq 0$, which contradicts assumption (5.2). Hence $\beta \geqslant \beta'$.

PROPOSITION 2: If α' and x are solutions of the quantity game, and β' and y are solutions of the price game, and if there is a correspondence between x and the active pure strategies in $M(\beta)$ and between y and the active pure strategies in $M(\alpha)$, then the quadruple (x, y, α' and β') are an $M-T$ solution of the system (1)–(5).

PROOF: By assumption (1), (2) and (5) are satisfied. The second part of the hypotheses proves that (4) and (5) are also satisfied.

PROPOSITION 3: If (x, y, α' and β') are $M-T$ solutions, then

$$\beta=\alpha\left(\frac{1+cp}{1+sp}\right), \qquad \text{where } p=\frac{xe}{xAy}.$$

ACKNOWLEDGEMENT

*I would like to thank O. Morgenstern and G. Thompson for their valuable criticisms of various points on an earlier draft of this paper.

REFERENCES

[1] Förstner, K., "Wirtschaftliches Wachstum bei vollständiger Konkurrenz," in: *Methods of Operations Research IV*, edited by Rufolf Henn (1968), pp. 1–257. (This important work of Karl Förstner is the most comprehensive discussion of the von Neumann model. I would like to mention, that the extremely clear exposition of the late German economist was very helpful to me on various points of this paper.)
[2] Förstner, K., op. cit., p. 75.
[3] Kaldor, N., "Alternative Theories of Distribution," Rev. of Econ. Studies (1955–56).
[4] Kemeny, T. G., O. Morgenstern, and G. L. Thompson, "A Generalization of the von Neumann Model of an Expanding Economy," Econometrica, **24**, 115–135 (1956).
[5] Kemeny, L. G., et al., op. cit., p. 122.
[6] Malinvaud, E., "Programmes d'Expansion et toux d'Interet," Econometrica, **27**, 215–227 (1959).
[7] Malinvaud, E., op. cit., p. 219–20.
[8] Malinvaud, E., op. cit., p. 227.
[9] Morgenstern, O., and L. G. Thompson, "Private and Public Consumption and Savings in the von Neumann Model of an Expanding Economy, "KYKLOS, Vol. **XX**, 387–409 (1967).
[10] Morishima, M., "Economic Expansion and the Interest Rate in Generalized von Neumann Models, "Econometrica, **28**, 352–363 (1960).
[11] Pasinetti, L., "Rate of Profit and Income Distribution in Relation to the Rate of Economic Growth," Rev. of Econ. Studies, **XXIX**, 267–279 (1962).
[12] Thompson, G. L., "On the Solution of a Game-Theoretic Problem," in *Linear Inequalities and Related Systems*, edited by H. W. Kuhn and A. W. Tucker (Princeton Univ. Press, Princeton, N.J., 1956), 280f.
[13] von Neumann, J., "A Model of General Economic Equilibrium," Rev. Econ. Stud. **13**, 1–9 (1945/46).

* * *

OPTIMAL POLICIES UNDER THE SHORTAGE PROBABILITY CRITERION FOR AN INVENTORY MODEL WITH UNKNOWN DEPENDENT DEMANDS*

Donald L. Iglehart

Stanford University

and

Stratton C. Jaquette

DECISION Studies Group

ABSTRACT

The principal innovation in this paper is the consideration of a new objective function for inventory models which we call the shortage probability criterion. Under this criterion we seek to minimize the total expected discounted cost of ordering subject to the probability that the stock level at the end of the period being less than some fixed quantity not exceed some prescribed number. For three different models we show that the minimum order policy is optimal. This result is then applied to a particular inventory model in which the demand distribution is not completely known. A Bayesian procedure is discussed for obtaining optimal policies.

1. INTRODUCTION

We consider a standard inventory model involving a single item with two new variations. First we shall consider a new objective function, which we term the shortage probability criterion, and describe the form of policies which are optimal under this criterion. Then we shall consider a particular inventory problem in which we have incomplete information on the dependent demands. We develop a Bayesian estimation procedure for the unknown demand parameter and shall find optimal ordering policy using this Bayesian estimate.

The inventory models which we study are not unusual. We shall assume that the state of the inventory is reviewed periodically (e.g., every week or month), and at these review points a decision is made to order new items. If new items are ordered, they arrive after a delay of λ periods ($\lambda = 0, 1, \ldots$) and are then added to the stock on hand. Between review points a stochastic demand, which we assume is non-negative, reduces the amount of stock on hand. We assume that excess demand is backlogged. The demands may depend on the amount of stock on hand and need not be identically distributed.

We shall impose the shortage probability criterion on the inventory model and seek ordering policies which are optimal relative to this criterion. For this criterion, assume that there is some level of stock which is just inadequate and denote this level by A. If the inventory level is less than or equal to A, then the inventory is short, while if the level is greater than A, the inventory is adequate. We require that the order at the beginning of each period be large enough so that the probability of being short at the end of each period does not exceed some prescribed probability. If orders are delivered

*This research was supported by the Office of Naval Research under contract Nonr–4457(00)(NR347–001).

after a delay of λ periods, an order in period i will not affect the probability of being short until period $i+\lambda$ and beyond. Thus when we are placing an order in period i, we are only concerned with shortages in periods $i+\lambda$, $i+\lambda+1$, ... We must always meet this probability constraint, but subject to this constraint we wish to minimize the total expected discounted cost of ordering. An ordering policy is optimal if it minimizes the expected discounted ordering cost over the period of interest subject to the shortage probability constraint.

We shall show that the minimum order policy is optimal for a variety of inventory models under the shortage probability criterion. The minimum order policy is the policy of ordering the minimum non-negative amount consistent with the constraint imposed on the probability of being short. We can express the probability of being short at the end of period $i+\lambda$ as a function of the amount we have on hand and order at the beginning of period i. If in period i we order the minimum quantity which meets the constraint for the shortage probability in period $i+\lambda$, then we are using the minimum order policy. If we are dealing with an n-period model, there is no probability constraint relevant to orders placed after period $n-\lambda$, and the order quantity for these periods under the minimum order policy is zero. The minimum order policy is a myopic policy since it is determined solely by the affect an order has on shortages during the period in which it arrives and not by its affect on the stock level in periods after it arrives.

An application of this theory concerns a particular inventory model with incompletely known demand distribution for which we seek an optimal ordering policy. For this model we shall assume that orders are delivered after a lag of one period ($\lambda=1$) and that demands during the period are binomially distributed with parameters n and p, where n is the amount of stock on hand at the beginning of the period and p is an unknown parameter satisfying $0<p<1$.

We shall develop a Bayesian procedure for estimating the parameter p. As is well known, the beta distribution is a particularly useful *a priori* distribution for p since the *a posteriori* distribution as we observe the inventory process over time is also a beta distribution. We shall show how this Bayesian estimate can be incorporated directly into the minimum order policy and that the policy is optimal for this inventory problem.

This paper is organized into three sections. In section 2 we prove that the minimum order policy is optimal under the shortage probability criterion for several inventory models. In section 3 we develop the Bayesian estimation scheme for the particular inventory model of interest and show how to incorporate it into an optimal ordering policy.

2. OPTIMALITY OF MINIMUM ORDER POLICY

An ordering policy is optimal for an n-period inventory problem if it minimizes the total expected discounted ordering cost for the n periods and satisfies the constraint on the probability of being short. We shall assume proportional ordering costs and for simplicity assume that the ordering cost per item is 1 and that it is incurred when the item is ordered. Assume a constant discount factor $\alpha(0\leq\alpha\leq 1)$. We shall denote by $C_n(x)$ the minimum expected discounted cost for an n-period model when the initial stock on hand plus on order is x. We shall find an ordering policy which attains this cost and which is consistent with the following shortage probability constraint.

The constraint on the probability of being short which is relevant to the order in period i can be expressed in the following form: the amount ordered, z, must satisfy

$$Pr\left\{x+z-\sum_{j=i}^{i+\lambda} D_j \leq A\right\}\leq\gamma,\ z\geq 0, \tag{1}$$

where x is the amount on hand plus on order at the beginning of period i just before the amount z is ordered, D_j is the stochastic demand experienced during period j, λ is the delivery lag, A is the shortage level for period $i+\lambda$, and γ is the shortage probability bound for period $i+\lambda$.

We shall find optimal ordering policies for three inventory models. Case 1 is a model with independent demands and arbitrary delivery lag. Case 2 is a model with demands which depend on the stock on hand at the beginning of the period and with immediate delivery of orders. Case 3 is the same as Case 2 except that deliveries arrive after a delay of 1 period. We shall not consider the more complicated case where $\lambda \geq 2$.

If we let y be the amount on hand plus on order after the current order is placed $(y=x+z)$, then we can express the probability of being short, the left most side of (1), as a simple function of x and y. We shall denote this probability by the function f. For Case 1, f depends only on y and

$$(2) \qquad f(y)=\int_0^\infty \cdots \int_0^\infty \phi(u_1) \cdots \phi(u_\lambda)\left[\int_{y-A-\sum_{j=1}^{\lambda} u_j}^{\infty} \phi(u)d\mu(u)\right]d\mu(u_1) \cdots d\mu(u_\lambda),$$

where $\phi(u)$ is the density with respect to μ of the demand distribution. We assume here and in what follows that μ is a σ-finite measure. For Case 2, f again depends only on y and

$$(3) \qquad f(y)=\int_{-\infty}^{A} \varphi(y,u)d\mu(u),$$

where $\varphi(y,u)$ is the density with respect to μ of the random variable representing the amount remaining on hand at the end of the period when y is on hand at the beginning. For Case 3, f depends on x and y and

$$(4) \qquad f(x,y)=\int_0^\infty \phi(x,u)\int_{-\infty}^{A} \varphi(y-u,v)d\mu(v)d\mu(u),$$

where $\phi(x,u)$ is the kernel of the demand distribution given that x is initially on hand. The functions ϕ and φ are related by $\phi(x,x-u)=\varphi(x,u)$.

We can now write the constraint on orders, (1), as an equivalent constraint on y involving the functions f

$$(5) \qquad f(y) \leq \gamma,\ y \geq x$$

or

$$(6) \qquad f(x,y) \leq \gamma,\ y \geq x.$$

We now state a result of Karlin's which we shall use repeatedly in the proofs which follow. For a proof of this lemma and a complete discussion of the concept of total positivity, we refer the reader to Karlin [1] and [2].

LEMMA 1: (Karlin [1].) *Suppose* $\phi(x, y)$ *is defined over the rectangle* $X \times Y$ *and*

$$\int_Y \phi(x,y)d\mu(y)=1 \qquad x \epsilon X.$$

Suppose also that ϕ *is totally positive of order* 2 (TP$_2$). *If* h(y) *is monotone, then*

$$g(x) \equiv \int_Y h(y)\phi(x,y)d\mu(y) \qquad x \epsilon X$$

is monotone in the same sense as h.

Requiring a density to be TP_2 is not a serious restriction since densities with monotone likelihood ratios are TP_2. Many common densities are TP_2, including the uniform distribution, the binomial distribution, and the exponential family of distributions.

As a direct application of Lemma 1 we can show that

$$F_x(a) \equiv \int_{-\infty}^{a} \phi(x, y) d\mu(y)$$

is nonincreasing as a function of x. Let $Y = X = (-\infty, +\infty)$, $h(y) = 1$ if $y \leq a$ and zero otherwise, and suppose ϕ is TP_2, then

$$F_x(a) = \int_Y h(y)\phi(x, y) d\mu(y)$$

is nonincreasing in x.

We now prove the optimality of the minimum order policy for our three cases of inventory models. Although we assume that demands are identically distributed, this is merely to simplify the exposition.

CASE 1.

LEMMA 2: Consider the inventory model of Case 1 and assume the demands have finite expected values, then f(y) *is nonincreasing in* y, *and for each* $\gamma (0 < \gamma \leq 1)$ *there is a* $\bar{y}$ *such that* $f(y) \leq \gamma$ *for* $y \geq \bar{y}$ *and* $f(y) > \gamma$ *for* $y < \bar{y}$.

PROOF: The monotonicity of $f(y)$ follows directly from the definition of $f(y)$ in (2). Since the expected value of demand is finite, clearly $f(y)$ goes to zero for large y and thus $\{y : f(y) \leq \gamma\}$ is non-empty. Define $\bar{y}$ by $\bar{y} \equiv \inf\{y : f(y) \leq \gamma\}$. Then $y > \bar{y}$ implies $f(y) \leq \gamma$ and $y < \bar{y}$ implies $f(y) > \gamma$, since f is monotone. Furthermore f is right continuous since the dependence on y is through a right continuous distribution function. Therefore, we know that $f(\bar{y}) \leq \gamma$ also.

We can easily characterize the minimum order policy for this model. The minimum order policy is to order up to the level $\bar{y}$; that is, order the quantity $z = [\bar{y} - x]^+$, where $[u]^+ = \max\ (u, 0)$. This is the minimum order since f is monotone and $\bar{y}$ just satisfies the probability constant. We can also replace the inequality (5) by

$$y \geq \bar{y} \vee x, \tag{5'}$$

where

$$u \vee v \equiv \max\ (u, v).$$

PROPOSITION 1: Suppose Lemma 2 holds, then the minimum order policy is optimal and $x + C_n(x)$ *is nondecreasing in* x.

PROOF: The proof follows that of Proposition 3 almost exactly, although the arguments are somewhat simpler for this case. We leave the details of this proof to the reader.

CASE 2.

LEMMA 3: Consider the inventory model of Case 2 and assume that φ *is* TP_2. *Then* f(y) *is nonincreasing in* y. *Suppose* $\{y : f(y) \leq \gamma\}$ *is non-empty for each* $\gamma (0 < \gamma \leq 1)$, *then let* $\bar{y} \equiv \inf\{y : f(y) \leq \gamma\}$. *If* $\varphi(y, u)$ *is right continuous in* y, *then* $y \geq \bar{y}$ *implies* $f(y) \leq \gamma$ *and* $y < \bar{y}$ *implies* $f(y) > \gamma$.

PROOF: By the definition of $f(y)$ in (3), $f(y)$ is the cumulative distribution of the amount left on hand at the end of the period, and $f(y) = F_y(A)$. By the results of Lemma 1, $F_y(A)$ and hence $f(y)$ is nonincreasing in y. The function f is right continuous since $\varphi(y, u)$ is right continuous in y by assumption. The rest of the proof follows that of Lemma 2.

The hypotheses of Lemma 3 are satisfied for many of the distributions of interest. For any distribution in the exponential family $\{y : f(y) \leq \gamma\}$ is non-empty. The continuity requirement is trivially satisfied for any discrete distribution. Distributions such as the uniform or binomial satisfy all the conditions of this lemma.

As in Case 1, the minimum order policy is to order up to $\bar{y}$, and we can replace (5) by (5′).

PROPOSITION 2: Suppose Lemma 3 holds and

$$\int_{-\infty}^{+\infty} u\varphi(y, u)\, d\mu(u) = ay + b \qquad a \leq 1.$$

Then the minimum order policy is optimal and $x + C_n(x)$ *is nondecreasing in* x.

PROOF: We shall use the dynamic programming recursion for $C_n(x)$

$$C_n(x) = \min_{y \geq \bar{y} \vee x} \{E[(y-x) + \alpha C_{n-1}(u)]\}, \tag{7}$$

where the expectation is relative to u, the amount left in stock at the end of the period. We shall define $C_0(x) \equiv 0$.

The proposition clearly holds for a 0-period model and can easily be shown true for a 1-period model. Assume the proposition true for an n-period model. To complete the proof we must show that it also holds for an $n+1$-period model.

From (7) we have

$$\begin{aligned} C_{n+1}(x) &= \min_{y \geq \bar{y} \vee x} \{(y-x) + \alpha \int \varphi(y, u) C_n(u)\, d\mu(u)\} \\ &= \min_{y \geq \bar{y} \vee x} \{-x - \alpha b + (1-\alpha a)y + \alpha \int \varphi(y, u)[u + C_n(u)]\, d\mu(u)\}, \end{aligned}$$

$(1-\alpha a)y$ is nondecreasing in y. By the induction hypothesis $u + C_n(u)$ is nondecreasing in u, hence by Lemma 1 $\int \varphi(y, u)[u + C_n(u)]\, d\mu(u)$ is nondecreasing in y. Thus the minimand is nondecreasing in y and the minimum is attained at the lower bound of y, namely at $y = \bar{y} \vee x$. Therefore, the optimal $n+1$-period ordering policy is the minimum order policy.

Thus $x + C_{n+1}(x)$ is given by

$$\begin{aligned} x + C_{n+1}(x) &= \bar{y} \vee x + \alpha \int \varphi(\bar{y} \vee x, u) C_n(u)\, d\mu(u) \\ &= (1-\alpha a)\bar{y} \vee x + \alpha \int \varphi(\bar{y} \vee x, u)[u + C_n(u)]\, d\mu(u) - \alpha b. \end{aligned}$$

Now by the same arguments this is nondecreasing in $\bar{y} \vee x$ and hence also in x since $\bar{y}$ is a constant.

CASE 3.

LEMMA 4: Consider the inventory model of Case 3 and assume that φ and ϕ are TP_2. *Then* $f(x, y)$ *is nonincreasing in* y *and nondecreasing in* x. *Suppose* $\{y : f(x, y) \leq \gamma\}$ *is non-empty for each* $\gamma (0 < \gamma \leq 1)$, *then* $\bar{y}(x) \equiv \inf\{y : f(x, y) \leq \gamma\}$ *is nondecreasing in* x. *If* $\varphi(y, u)$ *is right continuous in* y, *then* $y \geq \bar{y}(x)$ *implies* $f(x, y) \leq \gamma$, *and* $y < \bar{y}(x)$ *implies* $f(x, y) > \gamma$.

PROOF: The definition of $f(x, y)$ (4) contains the integral expression $\int_{-\infty}^{A} \varphi(y-u, v)d\mu(v)$. From the results of Lemma 1 this is nondecreasing in u and nonincreasing in y. By direct application of Lemma 1

$$f(x, y) = \int \phi(x, u) \int_{-\infty}^{A} \varphi(y-u, v)d\mu(v)d\mu(u)$$

is nondecreasing in x, and by a simpler argument $f(x, y)$ is nonincreasing in y. Suppose now that $x \leqslant x'$, then $f(x', y) \geqslant f(x, y)$ and in particular $f(x', \bar{y}(x)) \geqslant f(x, \bar{y}(x))$. For all $y < \bar{y}(x)$ we know $f(x, y) > \gamma$ and can conclude that $f(x', y) \geqslant f(x, y) > \gamma$. If $\bar{y}(x') < \bar{y}(x)$, choose y such that $\bar{y}(x') < y < \bar{y}(x)$, then $\gamma < f(x, y) \leqslant f(x', y) \leqslant \gamma$, a contradiction. Therefore $\bar{y}(x) \leqslant \bar{y}(x')$ and $\bar{y}(x)$ is nondecreasing in x. The rest of the proof follows as in Lemmas 2 and 3.

The assumption that φ and ϕ must both be TP_2 is somewhat restrictive. The uniform and binomial distributions are examples of distributions which do have this property.

We can make the same remarks about the simplicity of the minimum order rule here, and we can replace the inequality (6) by (6′)

$$y \geqslant \bar{y}(x) \vee x. \tag{6'}$$

PROPOSITION 3: Suppose Lemma 4 holds then the minimum order policy is optimal and $x + C_n(x)$ *is nondecreasing in* x.

PROOF: The proof will follow that for Proposition 2. We replace the dynamic programming recursion (7) by

$$C_{n+1}(x) = \min_{y \geqslant \bar{y}(x) \vee x} \left\{ (y-x) + \alpha \int \phi(x, u) C_n(y-u)d\mu(u) \right\}, \tag{8}$$

where the expectation is relative to the demand random variable.

Since the proposition trivially holds for 0- and 1-period problems, we shall only prove the induction step. From (8) we obtain

$$C_{n+1}(x) = \min_{y \geqslant \bar{y}(x) \vee x} \left\{ -x + \alpha \int u\phi(x, u)d\mu(u) + (1-\alpha)y + \alpha \int \phi(x, u)[C_n(y-u) + (y-u)]d\mu(u) \right\}.$$

The minimand is nondecreasing in y since $\alpha < 1$ and since $x + C_n(x)$ is nondecreasing by assumption. The minimum value is therefore attained at $\bar{y}(x) \vee x$, and the minimum order policy is optimal for the $n+1$-period model. We can also examine $x + C_{n+1}(x)$ which is

$$\begin{aligned} x + C_{n+1}(x) &= \bar{y}(x) \vee x + \alpha \int \phi(x, u) C_n(\bar{y}(x)\, x - u)d\mu(u) \\ &= (1-\alpha)(\bar{y}(x) \vee x) + \alpha \int \phi(x, u)[C_n(\bar{y}(x)\, x - u) \\ &\qquad + (\bar{y}(x) \vee x - u)]d\mu(u) + \alpha \int u\phi(x, u)d\mu(u). \end{aligned}$$

Since ϕ is TP_2 and the function $h(u) = u$ is increasing, by Lemma 1 $\int u\phi(x, u)d\mu(u)$ is nondecreasing in x. The rest of the terms are nondecreasing in $\bar{y}(x) \vee x$ by previous arguments. Since $\bar{y}(x)$ is nondecreasing in x we can conclude that the entire expression is nondecreasing in x and that the proposition holds for all n.

Propositions 1, 2, and 3 can easily be generalized to cover nonstationary inventory models where we allow changing demand distributions, shortage levels, and shortage probabilities. We need only redefine the f's and $\bar{y}$'s as follows

$$(2') \qquad f_i(y) = \int_0^\infty \cdots \int_0^\infty \phi_i(u_1) \cdots \phi_{i+\lambda-1}(u_\lambda) \left[\int_{y-A_{i+\lambda}-\sum_{j=1}^{\lambda} u_j}^\infty \phi_{i+\lambda}(u) d\mu(u) \right] d\mu(u_1) \cdots d\mu(u_\lambda).$$

$$(3') \qquad f_i(y) = \int_{-\infty}^{A_i} \varphi_i(y, u) d\mu(u).$$

$$(4') \qquad f_i(x, y) = \int_0^\infty \phi_i(x, u) \int_{-\infty}^{A_{i+1}} \varphi_{i+1}(y-u, \omega) d\mu(\omega) d\mu(u)$$

$$\bar{y}_i \equiv \inf \{y : f_i(y) \leq \gamma_{i+\lambda}\}$$

$$\bar{y}_i(x) \equiv \inf \{y : f_i(x, y) \leq \gamma_{i+1}\}.$$

The minimum order policy is to order up to $\bar{y}_i$ in period i.

Propositions 1, 2, and 3 show that the minimum order policy is dynamically optimal for nonstationary inventory models when demands are independent and the delivery lag is arbitrary or when demands are state dependent and $\lambda = 0$ or 1.

3. AN INVENTORY MODEL WITH BAYESIAN ESTIMATION OF DEPENDENT DEMAND

Consider now an inventory model with an incompletely known demand distribution for which we shall develop a Bayesian estimator of demands. Consider a Case 3 inventory model with a one period delivery lag and with demands binomially distributed with parameters x and p. Let x be the number of items on hand at the beginning of the period and $p (0 < p < 1)$ be an unknown parameter of the binomial distribution. This particular inventory model can arise if the demand is due to deterioration or failure of the items, and if the time to failure is exponentially distributed with unknown mean. The total number of failures, or demand, is then binomially distributed with unknown parameter p.

Let x_i be the stock on hand at the start of the ith period and d_i be the demand during that period, $i = 1, 2, \ldots, n$. We know that the pair

$$\left(\sum_{i=1}^n d_i, \sum_{i=1}^n x_i \right)$$

is a sufficient statistic for p. The maximum likelihood estimator, $\hat{p}$, of p is a function of the sufficient statistic and is also unbiased;

$$\hat{p} = \sum_{i=1}^n d_i \bigg/ \sum_{i=1}^n x_i.$$

We want a simple Bayesian estimator which depends on the sufficient statistic, for then we know that the *a posteriori* Bayesian estimate converges in law to the true value and has many of the same desirable properties of the maximum likelihood estimator.

We choose a beta distribution as the *a priori* distribution for the parameter p and show, as is well known, that the *a posteriori* distribution is again a beta distribution. Suppose we take the *a priori*

distribution of p to be

$$f(u|\alpha,\beta)=(\beta+1)\binom{\beta}{\alpha}u^{\alpha}(1-u)^{\beta-\alpha},\ 0\leq u\leq 1,$$

where α and β ($\alpha\leq\beta$) are the two parameters specifying the beta distribution. For the purposes of this paper, we shall assume that α and β are integers, although in general they need only be real. If we then observe a period in which demand is d from an initial stock of x, we can find the *a posteriori* distribution of p which is given by

$$\begin{aligned}f(u|\alpha,\beta;d,x)&=\frac{f(u|\alpha,\beta)\binom{x}{d}u^{d}(1-u)^{x-d}}{\int_0^1 f(v|\alpha,\beta)\binom{x}{d}v^{d}(1-v)^{x-d}dv}\\&=(\beta+x+1)\binom{\beta+x}{\alpha+d}u^{\alpha+d}(1-u)^{(\beta+x)-(\alpha+d)}\\&=f(u|\alpha+d,\beta+x).\end{aligned}$$

This *a posteriori* distribution then becomes the new *a priori* distribution of p before we observe the next period. Since the form of the distribution of p is preserved, if we start with an initial *a priori* distribution $f(u|\alpha_0,\beta_0)$, the *a posteriori* distribution of p after using the Bayesian estimation scheme for n periods is just $f(u|\alpha_n,\beta_n)$, where

$$\alpha_n=\alpha_{n-1}+d_n=\alpha_0+\sum_{i=1}^{n}d_i$$

$$\beta_n=\beta_{n-1}+x_n=\beta_0+\sum_{i=1}^{n}x_i.$$

We can calculate the distribution of demands when the *a priori* distribution is specific by α and β

$$\begin{aligned}\Pr\{D=d|x,\alpha,\beta\}&=\int_0^1 f(u|\alpha,\beta)\binom{x}{d}u^{d}(1-u)^{x-d}du\\&=\binom{x}{d}(\beta+1)\binom{\beta}{\alpha}\Big/(\beta+x+1)\binom{\beta+x}{\alpha+d},\end{aligned}\tag{9}$$

which approaches the true distribution of demand as the number of observations increases, namely,

$$Pr\{D=d|x,p\}=\binom{x}{d}p^{d}(1-p)^{x-d}.$$

We can now easily describe the optimal ordering policy for this inventory model when we use this Bayesian estimate of the unknown parameter. We give the result in Proposition 4.

PROPOSITION 4: For the above inventory problem the minimum order policy is optimal. The optimal policy is to order up to a level $\bar{y}(x)$, where $\bar{y}(x)$ is the minimum value of y satisfying

$$\sum_{k=0}^{A}\sum_{i=0}^{x}\binom{x}{i}\binom{y-i}{k}(\beta+1)\binom{\beta}{\alpha}/\left((\beta+x+y-i+1)\binom{\beta+x+y-i}{\alpha+y-k}\right)\leq\gamma.$$

PROOF: Since we are dealing with discrete random variables, we need only verify that the distributions of D in (9) and of $x-D$ are TP_2. Since these distributions are TP_2, we know from Proposition 2 that the minimum order policy is optimal.

The optimal policy for this problem approaches the optimal policy for the same inventory problem where the demand distribution is completely known. For completeness we present the optimal policy for this case below.

PROPOSITION 5: For the above inventory problem with the demand distribution known, the minimum order policy is optimal, and $\bar{y}(x)$ is the minimum value of y satisfying

$$\sum_{k=0}^{A} \sum_{i=0}^{x} \binom{x}{i}\binom{y-i}{k} p^{y-k}(1-p)^{x-i+k} \leq \gamma.$$

The significance of Proposition 4 is that we can characterize the policy which is dynamically optimal for an inventory problem where we must estimate the unknown dependent demands.

REFERENCES

[1] Karlin, S., "Total Positivity and Convexity Preserving Transformations," *Proc. of Sym. in Pure Math.*, 7, 329–347 (1963).
[2] Karlin, S., *Total Positivity* (Stanford University Press, Stanford, Calif., 1967).

MULTI-CLASS INVENTORY MODELS WITH DEMAND A FUNCTION OF INVENTORY LEVEL*

Donald L. Iglehart
Stanford University

and

Stratton C. Jaquette
DECISION Studies Group

ABSTRACT

This paper considers the problem of maintaining an inventory of an item which can deteriorate and become useless. A periodic review procedure is used and new items ordered may experience a time lag in delivery. Items are considered to deteriorate through one or two states before becoming useless. Thus the deterioration process in each period plays the role of the usual demand process and is a function of the inventory level at the beginning of each period. For the case of no time lag in delivery, one stage deterioration, and either binomial or uniform deterioration, optimal ordering policies are obtained for the n-period dynamic model with the standard cost structure. (For the shortage probability criterion see the other paper by Iglehart and Jaquette, in this issue.) These policies are of the single critical number type. For more complicated models suboptimal policies of this same type are found.

1. INTRODUCTION

We shall consider inventory models in which a single item can exist in one of $k+1$ classes. These classes can be thought of as a classification of items according to some physical characteristic such as age, condition, or location. For this exposition we shall assume that items are classified according to their condition, with class 1 denoting items in perfect (new) condition while items in class $k+1$ are worthless (permanently failed). The intermediate classes $(2, 3, \ldots, k)$ represent items in increasingly worse condition. A periodic review procedure will be used. Between review points items in class $1, 2, \ldots, k$ are permitted to deteriorate to at most the next higher class (next worst condition) according to a given stochastic law. Items in class $k+1$ remain there. At review points new items of class 1 can be ordered and they are assumed to be delivered after a lag of $\lambda(\lambda=0, 1, \ldots)$ periods. A cost structure is superimposed on this stochastic model and either optimal or suboptimal policies are sought.

This model differs from existing inventory models principally in the assumption that demands (i.e., the number of items deteriorating in each class during a period) depend on the number of items on-hand in each class at the beginning of the period. For example, if items are in class j at the beginning of a period, the number deteriorating during the period might have a binomial distribution with parameters n and $p_j(0<p_j<1)$. Most stochastic inventory models assume that demands in successive periods constitute a sequence of independent, identically distributed random variables which are not a function of inventory levels.

*This research was supported by the Office of Naval Research under contract Nonr-4457(00)(NR347-001).
We are grateful to Dr. Stuart A. Bessler for many helpful discussions in the formulation of these models.

We have described the basic model in greater generality than we have been able to solve. For the case $k=1$ and $\lambda=0$, we are able to obtain the complete optimal dynamic solution for an n-period problem. In the cases where $k=1$ or 2 and $\lambda=0$ or 1, we have assumed the form of the ordering policy and then found the stationary distribution of the induced Markov chain for binomial and uniform demands. In these cases we are then able to sub-optimize by selecting the best policy of the given form.

This paper is organized into the following sections. Section 2 deals with some preliminaries on convexity preserving transformations. In section 3 the optimal dynamic policy for $k=1$, $\lambda=0$ is obtained. Sections 4 and 5 are concerned with the stationary analysis of the models $k=1$, $\lambda=0$ or 1 and $k=2$, $\lambda=0$ or 1, respectively. Finally, section 6 treats the case $k=2$, $\lambda=0$ or 1 for a more complicated form of ordering policy.

2. CONVEXITY PRESERVING TRANSFORMATIONS

In our proofs of optimality in succeeding sections, we shall make strong use of the following class of functions. Let X and Y be two subsets of the real line and $K(x, y)$ be a real-valued function on $X\times Y$. Then K is said to be *totally positive of order r* (abbreviated TP_r) if for all m $(1 \leq m \leq r)$ and for all

$$x_1 < x_2 < \dots < x_m;\ y_1 < y_2 < \dots < y_m\ (x_i \epsilon X,\ y_j \epsilon Y),$$

we have the inequalities

$$K\begin{pmatrix} x_1, x_2, \dots, x_m \\ y_1, y_2, \dots, y_m \end{pmatrix} = \begin{vmatrix} K(x_1, y_1) K(x_1, y_2) \dots K(x_1, y_m) \\ K(x_2, y_1) K(x_2, y_2) \dots K(x_2, y_m) \\ \vdots \qquad \vdots \qquad \vdots \\ K(x_m, y_1) K(x_m, y_2) \dots K(x_m, y_m) \end{vmatrix} \geq 0.$$

We shall be interested in sets X and Y which are either of the form $\{0, 1, 2, \dots, N\}$ or $(0, N)$. If f is a convex function, then Proposition 1 gives a sufficient condition for g to be convex, where g is defined as

$$g(x) = \int_Y K(x, y) f(y)\, d\mu(y) \qquad x \epsilon X,$$

with μ a σ-finite measure.

PROPOSITION 1: (Karlin [3].) *Let X and Y be open intervals of the real line or an interval of integers (i.e., a consecutive set of integers). Assume that*

$$\text{(a)} \int_Y K(x, y)\, d\mu(y) = 1 \qquad x \epsilon X,$$

$$\text{(b)} \int_Y yK(x, y)\, d\mu(y) = ax + b \qquad x \epsilon X, \qquad a > 0,$$

and

(c) $K(x, y)$ *is* TP_3.

If $f(y)$ is convex (concave), then $g(x)$ is convex (concave).

The examples of greatest interest for our inventory models are the following.

(i) $K(n, j) = \binom{n}{j} p^j (1-p)^{n-j}$, $n=0, 1, 2, \dots$; $j=0, 1, \dots, n$; $0 < p < 1$; and μ is counting measure; *cf.* Karlin and Proschan [2]. This is, of course, the binomial distribution whose mean is np.

(ii) $K(n, j) = (n+1)^{-1}$, $j=0, 1, \dots, n$, $n=1, 2, \dots$, and μ is counting measure. This is the discrete uniform distribution with mean $n/2$. The continuous uniform distribution also satisfies the condi-

tions of Proposition 1. For a comprehensive discussion of the subject of totally positive functions and their applications the reader should consult Karlin [4].

The other result we shall need on preserving convex convex functions is given in Proposition 2. This result is very useful for certain functional equations which arise in dynamic programming. A proof of the result is given in Iglehart [1].

PROPOSITION 2: Let

$$R(\mathbf{x}) = \min_{y \epsilon D(\mathbf{x})} \{M(\mathbf{x}, \mathbf{y})\} = M[\mathbf{x}, \mathbf{y}_0(\mathbf{x})]y_0(\mathbf{x}) \epsilon D, \ x \epsilon C,$$

where $\mathbf{x}$ *and* $\mathbf{y}$ *are* k*-dimensional vectors,* $R(\cdot)$ *and* $M(\cdot, \cdot)$ *are real-valued functions, and* $D(\mathbf{x})$ *is some domain in* R^k *which depends on* x*. If* $M(\mathbf{x}, \mathbf{y})$ *is a convex function of the* $2k$*-dimensional vector* $\mathbf{z} = (\mathbf{x}, \mathbf{y})$, C *is convex, and the set* $\{(\mathbf{x}, \mathbf{y}) : \mathbf{y} \epsilon D(\mathbf{x})\}$ *is convex, then* $R(\mathbf{x})$ *is convex in* $\mathbf{x}$.

3. OPTIMAL DYNAMIC POLICY FOR THE MODEL $k=1, \lambda=0$

Consider a two-class model ($k=1$) in which items in perfect condition may undergo a stochastic transition to the worthless condition between review points. All items ordered are delivered instantaneously ($\lambda=0$) and the demand (number of items deteriorating) has either the binomial or uniform distribution (examples (i) and (ii) of section 2). Three costs will be charged each period which influence our ordering decisions. A convex increasing ordering cost $c(z)$ is charged when z items are ordered with $c(0)=0$. In addition, we shall charge holding and penalty costs based on the inventory level at some specified point of the review period. Let h and p be convex, increasing holding and shortage costs, respectively, which vanish for non-positive arguments. Then, for example, if we charge both of these costs at the end of the review period where x is the end of period inventory level, the holding and shortage costs for that period are $l(x) = h([x-a]^+) + p([a-x]^+)$. Here a is a specified inventory level and $[y]^+ = \max(y, 0)$. Clearly, l is convex. A second alternative would be, for example, to charge the holding cost at the beginning of the period and the shortage cost at the end of the period. If y is the inventory level at the beginning of the period after ordering, the one-period holding and shortage cost would be $l(y; x) = h([y-a]^+) + p([a-x]^+)$. Let $D(y)$ be the demand in the period when the inventory after ordering is y. The demand $D(y)$ can be thought of as the number of items (out of y) making the transition to the worthless condition. Then for the first alternative above the expected one-period holding and shortage cost would be $L(y) = E[l(y-D(y))]$, which is convex by Proposition 1. For the second alternative $L(y) = h([y-a]^+) + E[p([a-y+D(y)]^+)]$ which is also convex in y. In general for the theory which follows all we need is that $L(y)$, the expected one-period holding and shortage cost as a function of the inventory level y after ordering, be convex. We include a discount factor α, $0 < \alpha \leq 1$.

Let x be the initial inventory level in an n-period model and $C_n(x)$ the corresponding optimal cost. Then the functional equation of dynamic programming becomes

$$C_n(x) = \min_{y \geq x} \{c(y-x) + L(y) + \alpha E[C_{n-1}(y-D(y))]\} \qquad x \geq 0,$$

where $C_0(x) \equiv 0$. The following proposition gives the optimal ordering policy for this model.

PROPOSITION 3: Under the conditions stated above, we have

(a) *the optimal ordering policy is characterized by a integer-valued function* $\bar{y}_n(x)$ *and an integer* $\bar{x}_n$*; if* $x < \bar{x}_n$*, order up to* $\bar{y}_n(x)$*, and if* $x \geq \bar{x}_n$ *do not order;*

(b) $C_n(x)$ *is convex.*

PROOF: As usual the proof is by induction on n. For $n=1$ let $\bar{y}_1(x) \geqslant x$ minimize the convex function $c(y-x)+L(y)$ in the interval $[x, \infty)$ and $\bar{x}_1$ be the smallest integer such that

$$c(1)+L(x+1)-L(x) \geqslant 0.$$

Then the optimal policy is clearly given by (a) and C_1 is convex by Proposition 2. Assume that the result is true for $n=N-1$. Since

$$E[C_{N-1}(y-D(y))]=\int_0^y C_{N-1}(\xi)K(y, \xi)d\mu(\xi),$$

after a change of variables, $E[C_{N-1}(y-D(y))]$ is convex in y by Proposition 1. Here we have used the fact that, in the notation of examples (i) and (ii), $n-j$ is also either binomial or uniform according to what j is. Hence C_N is again convex by Proposition 2. Let $\bar{y}_N(x)$ minimize the convex function $c(y-x)+L(y)+E[C_{N-1}(y-D(y))]$ in the interval $[x, \infty)$ and $\bar{x}_N$ be the smallest integer such that $c(1)+L(x+1)-L(x)+E[C_{N-1}(x+1-D(x+1))]-E[C_{N-1}(x-D(x))] \geqslant 0$. With these definitions the optimal policy of (a) is clearly established and the proof of the proposition completed.

It is also possible to show that the $\bar{y}_n(x)$ and $\bar{x}_n$ are nondecreasing in n and that the infinite horizon optimal policy has the same form. These results can be established using the methods of [1] and are left to the reader.

For the special case of a linear ordering cost $c(x)=c \cdot x$ with $c>0$ the following result is easily proved.

COROLLARY 1: If $c(x)=c \cdot x$, *then* $\bar{y}_n(x)=\bar{x}_n$.

If in addition to the linear ordering cost we assume that any items remaining at the end of the last period can be sold for αc, then $\bar{x}_n \equiv \bar{x}$ where $\bar{x}$ is the minimum of the convex function

$$c(1-\alpha)y+E[D(y)]+L(y).$$

This result can be obtained either by including a salvage cost in the above dynamic programming formulation or by following the method of Veinott [5]. It has the appealing feature of only requiring one computation to obtain the n-period optimal policy for all values of n.

4. STATIONARY ANALYSIS OF THE MODEL $k=1$ AND $\lambda=0$ OR 1

We consider in this section the stationary analysis of the two-class ($k=1$) model with time lag in delivery $\lambda=0$ or 1 and a single critical number policy (order up to a fixed level). The demand distributions considered are either the binomial or uniform. Our one-period holding and shortage cost $l(y)=h \cdot [y-a]^+ + p \cdot [a-y]^+$ and the ordering cost is linear, $c \cdot z$. We shall first obtain the stationary distribution of the inventory level before ordering and then optimize the stationary expected cost with respect to the single critical number.

We begin with the case $\lambda=0$ which is rather trivial. Let s be the single critical number which we order up to, X_n be the inventory level before ordering at the beginning of the nth period, and $D(s)$ the demand in the nth period. Then the transition rule is

$$X_{n+1}=s-D(s).$$

Since the right-hand side is independent of X_n, the stationary distribution of the Markov chain $\{X_n : n=1, 2, \ldots\}$ is simply the distribution of $s-D(s)$. Hence

$$P[X_{n+1}=i]=P[D_n(s)=s-i]=\begin{cases} (s+1)^{-1}, & \text{uniform demand} \\ \binom{s}{s-i} p_1^{s-i}(1-p_1)^i, & \text{binomial demand,} \end{cases}$$

where p_1 is the probability of a single item failing in one period.

The stationary one-period expected cost using the policy characterized by s is

(1) $$C(s)=E\{c\cdot D(s)+h\cdot[s-D(s)-a]^{+}+p\cdot[a-s+D(s)]^{+}\}.$$

For convenience, we assume $c<p$, otherwise the optimal $s^*=0$. The optimal s^* is easy to calculate and we list the results as Proposition 4. We let $(h+p)/(h+c)\equiv\rho$.

PROPOSITION 4: For the model $k=1$, $\lambda=0$ and the cost function (1) the optimal values of s are

(i) *s^* is the largest integer s satisfying*

$$1\leq\rho\sum_{i=0}^{a-1}\binom{s}{i}p_1^{s-i}(1-p_1)^{i}$$

for a binomial demand, and

(ii) $s^*=\left[-\dfrac{3}{2}+(a^2\rho+a\rho+1/4)^{1/2}\right]$ *for a uniform demand.*

We turn now to the case of a one-period delivery lag ($\lambda=1$). Now the transition rule is

$$X_{n+1}=s-D_n(X_n).$$

The transition probability matrix $\{p_{ij}\}$ for the Markov chain $\{X_n\}$ with uniform demand is given by

$$p_{ij}=\begin{cases}(i+1)^{-1} & j=s-i,\ .\ .\ .\ s\\ 0 & \text{otherwise.}\end{cases}$$

The stationary distribution is found to be triangular; i.e.,

$$\pi_j=\frac{2}{(s+1)(s+2)}(j+1)\qquad j=0,1,\ldots,s.$$

The mean,

$$\sum_{j=0}^{s}j\pi_j,$$

is equal to (2/3)s. For the binomial demand the transition matrix has elements

$$p_{ij}=\binom{i}{s-j}p_1^{s-j}(1-p_1)^{i-(s-j)}\qquad j=s-i,\ldots,s,$$

and the stationary distribution is also binomial; i.e.,

$$\pi_j=\binom{s}{j}\left(\frac{1}{1+p_1}\right)^{j}\left(\frac{p_1}{1+p_1}\right)^{s-j}\qquad j=0,1,\ldots,s$$

with mean $s/(1+p_1)$.

In this case the one-period stationary expected cost using policy s is

(2) $$C(s)=E\{c\cdot D(X)+h\cdot[X-D(X)-a]^{+}+p\cdot[a-X+D(X)]^{+}\},$$

where X is a random variable whose distribution is the stationary distribution of the Markov chain $\{X_n\}$ and $D(X)$ is the associated demand. Let $\rho=(h+p)/(h+2p_1c/(1-p_1))$. Then the optimal policies are given by

PROPOSITION 5: For the model $k=1$, $\lambda=1$ and the cost function (2) the optimal values of s are

(i) *$s^*=0$ if $\rho\leq1$, and otherwise s^* is the largest integer s satisfying*

$$1\leq\rho\sum_{i=0}^{a-1}\binom{s}{i}\left(\frac{1-p_1}{1+p_1}\right)^{i}\left(\frac{2p_1}{1+p_1}\right)^{s-i}$$

for a binomial demand; and

(ii) $s^* = 0$ *if* $c \geqslant p$, *and otherwise* s^* *is obtained from the roots of a* 3rd *order polynomial for a uniform demand.**

5. STATIONARY ANALYSIS FOR THE MODEL $k=2$ AND $\lambda=0$ OR 1

In this section we extend the results of section 4 to the three-class ($k=2$) model with either uniform or binomial demand distributions and $\lambda=0$ or 1. Again we shall follow a single critical number ordering rule with respect to the total amount of inventory. Let W_n be the total inventory level of the items in classes 1 and 2 and Y_n the level in class 2 before ordering in the nth period. Then the transition rule in the case $\lambda=0$ becomes

$$W_{n+1} = s - E_n(Y_n)$$
$$Y_{n+1} = Y_n - E_n(Y_n) + D_n(s - Y_n),$$

where D_n is the demand for items in class 1 in the nth period and E_n is the corresponding demand for class 2. Clearly, $\{Y_n\}$ is a one-dimensional Markov chain. For the case where D_n and E_n are independent and uniformly distributed the stationary distribution for $\{Y_n\}$ is given by

$$\pi_j = \frac{6}{(s+1)(s+2)(s+3)} \cdot (j+1)(s-j+1) \qquad j=0, 1, \ldots, s.$$

If D_n and E_n are independent binomials with parameters p_1 and p_2, respectively, then

$$\pi_j = \binom{s}{j} \left(\frac{p_1}{p_1+p_2}\right)^j \left(\frac{p_2}{p_1+p_2}\right)^{s-j} \qquad j=0, 1, \ldots, s.$$

The one-period stationary cost using policy s is

$$C(s) = E\{c \cdot E(Y) + h \cdot [W - E(Y) - a]^+ + p \cdot [a - W + E(Y)]^+\}, \tag{3}$$

where W and Y are the stationary random variables of the Markov chain $\{(W_n, Y_n)\}$ and $E(Y)$ is the associated demand. Let

$$\rho \equiv (h+p) \Big/ \left(h + \frac{cp_1p_2(2p_1+p_2-p_1p_2)}{(p_1+p_2)(p_1+p_2-p_1p_2)} \right).$$

Then the optimal policies are given by

PROPOSITION 6: For the model $k=2$, $\lambda=0$ *and the cost function* (3) *the optimal values of* s *are*

(i) $s^*=0$ *if* $c \geqslant p(p_1+p_2-p_1p_2)/p_1p_2$, *and otherwise* s^* *is the largest integer* s *satisfying*

$$1 \leqslant \rho \sum_{i=0}^{a-1} \binom{s}{i} \left(\frac{p_1+p_2-p_1p_2}{2p_1+p_2-p_1p_2}\right)^i \left(\frac{p_1}{2p_1+p_2-p_1p_2}\right)^{s-i}$$

for a binomial demand; and

(ii) $s^*=0$ *if* $c \geqslant 3p$, *and otherwise* s^* *is obtained from the roots of a* 4th *order polynomial for a uniform demand.*

Turning now to the case $\lambda=1$ the transition rule becomes

$$W_{n+1} = s - E_n(Y_n)$$
$$Y_{n+1} = Y_n - E_n(Y_n) + D_n(W_n - Y_n).$$

*This polynomial and the specific value of s^* have been calculated; however, they are so complicated that we have elected to omit them. A similar remark holds for the polynomials which arise in Propositions 6 and 7.

In this case, the pair (W_n, Y_n) form a Markov chain. If we let π_{ij} be the stationary distribution of this Markov chain, then when E_n and D_n are uniform

$$\pi_{ij}=\frac{24}{(s+1)(s+2)(s+3)(s+4)}(j+1)(i-j+1)\ j\leq i=0,1,\ldots,s.$$

The marginal distribution for W, π_i, is given by

$$\pi_i=\frac{4}{(s+1)(s+2)(s+3)(s+4)}(k+1)(k+2)(k+3)\ i=0,1,\ldots,s.$$

When E_n and D_n are binomial we have

$$\pi_{ij}=\binom{s}{i}\left(\frac{p_1+p_2}{p_1+p_2+p_1p_2}\right)^i\left(\frac{p_1p_2}{p_1+p+p_1p_2}\right)^{s-i}\binom{i}{j}\left(\frac{p_1}{p_1+p_2}\right)^j\left(\frac{p_2}{p_1+p_2}\right)^{i-j},$$

for $j\leq i=0,1,\ldots,s$. The marginal distribution for W is

$$\pi_i=\binom{s}{i}\left(\frac{p_1+p_2}{p_1+p_2+p_1p_2}\right)^i\left(\frac{p_1p_2}{p_1+p_2+p_1p_2}\right)^{s-i}\ i=0,\ldots,s.$$

For this case the one-period stationary cost is also given by (3). Now we let

$$\rho\equiv(h+p)/\left(h+\frac{cp_1p_2}{p_1+p_2+p_1p_2}\right).$$

Then we have

PROPOSITION 7. For the model $k=2$, $\lambda=1$ and the cost function (3) the optimal values of s are

(i) $$s^*=0\ \mathit{if}\ c\geq p\left(\frac{p_1+p_2+p_1p_2}{p_1p_2}\right),$$

and otherwise s^* is the largest integer satisfying

$$1\leq\rho\sum_{i=0}^{a-1}\binom{s}{i}\left(\frac{p_1+p_2-p_1p_2}{p_1+p_2+p_1p_2}\right)^i\left(\frac{2p_1p_2}{p_1+p_2+p_1p_2}\right)^{s-i}$$

for a binomial demand; and

(ii) *s^* is obtained from the roots of a* 5th *order polynomial for a uniform demand.*

6. A MORE COMPLICATED ORDERING RULE FOR THE CASE $k=2$, $\lambda=1$

In the last section we analyzed the two-class model when the ordering rule was a single critical number for the total inventory level. For the two-class model, we have not been able to characterize the optimal ordering policy; however, it seems unlikely that the optimal policy is of the form stated above since the distribution of items between the two classes would seem to be relevant. With this motivation we consider the ordering policy based on two parameters s and b, $0<b<1$. In each period we order $s-W_n+[bY_n]$, where $[x]$ is the integer part of x. For the case $\lambda=0$ the transition rule is

(4) $$W_{n+1}=s+[bY_n]-E_n(Y_n)$$

(5) $$Y_{n+1}=Y_n+D_n(s+[bY_n]-Y_n)-E_n(Y_n).$$

For the case $\lambda=1$,

(6) $$W_{n+1}=s+[bY_n]-E_n(Y_n)$$

(7) $$Y_{n+1}=Y_n+D_n(W_n-Y_n)-E_n(Y_n).$$

In both cases the pairs (W_n, Y_n) form a Markov chain; however, the stationary distributions satisfy a very complicated system of linear equations which we have not been able to solve. It is possible, however, to calculate the first few moments of this stationary distribution. If our functions are linear or quadratic, these stationary moments do allow us to find the stationary one-period cost.

To calculate, for example, the stationary expected value of W and Y in a heuristic manner one would take expected values in the pairs (4), (5) or (6), (7), remove the subscripts n, and solve the resulting algebraic equations for $E[W]$ and $E[Y]$. For specific cases this is easily done and we leave it to the interested reader.

REFERENCES

[1] Iglehart, D. L., "Capital Accumulation and Production for the Firm: Optimal Dynamic Policies," Management Sci. *12*, 193–205 (1965).

[2] Karlin, S. and F. Proschan, "Polya Type Distributions of Convolutions," Ann. Math. Stat. *31*, 721–736 (1960).

[3] Karlin, S., "Total Positivity and Convexity Preserving Transformations," Proc. of Sym. in Pure Math., 7, 329–347 (1963).

[4] Karlin, S., *Total Positivity* (Stanford Univ. Press, Stanford, Calif., 1967).

[5] Veinott, A. F., "Optimal Policy in a Dynamic, Single Product, Nonstationary Inventory Model with Several Demand Classes," Operations Res. *13*, 761–778 (1965).

* * *

LOCATING FACILITIES IN THREE-DIMENSIONAL SPACE BY CONVEX PROGRAMMING

Robert F. Love*

The University of Wisconsin
Madison, Wisconsin

ABSTRACT

We consider the problem of simultaneously locating any number of facilities in three-dimensional Euclidean space. The criterion to be satisfied is that of minimizing the total cost of some activity between the facilities to be located and any number of fixed locations. Any amount of activity may be present between any pair of the facilities themselves. The total cost is assumed to be a linear function of the inter-facility and facility-to-fixed locations distances. Since the total cost function for this problem is convex, a unique optimal solution exists. Certain discontinuities are shown to exist in the derivatives of the total cost function which previously has prevented the successful use of gradient computing methods for locating optimal solutions. This article demonstrates the use of a created function which possesses all the necessary properties for ensuring the convergence of first order gradient techniques and is itself uniformly convergent to the actual objective function. Use of the fitted function and the dual problem in the case of constrained problems enables solutions to be determined within any predetermined degree of accuracy.

Some computation results are given for both constrained and unconstrained problems.

I. INTRODUCTION

We consider the problem of finding the locations of m facilities in three-dimensional Euclidean space which are optimal with respect to p other fixed locations.

Let the coordinates of the ith unlocated facility be given by $x_i = (x_{1i}, x_{2i}, x_{3i})$, and the coordinates of the jth fixed location by $a_j = (a_{1j}, a_{2j}, a_{3j})$, $i = 1, \ldots, m$; $j = 1, \ldots, p$. The total cost function is given by

$$(1) \qquad C(x) = \sum_{j=1}^{p} \sum_{i=1}^{m} c_{ij}[(a_{1j} - x_{1i})^2 + (a_{2j} - x_{2i})^2 + (a_{3j} - x_{3i})^2]^{1/2}$$
$$+ \sum_{i=1}^{m} \sum_{t=1}^{m} \bar{c}_{it}[(x_{1i} - x_{1t})^2 + (x_{2i} - x_{2t})^2 + (x_{3i} - x_{3t})^2]^{1/2},$$

where $x \equiv (x_{11}, x_{21}, x_{31}, \ldots, x_{ij}, \ldots, x_{3m})$ and $\bar{c}_{it} = 0$, when $i = t$.

Each of the c_{ij} values in Eq. (1) is a constant representing cost per unit distance between the jth fixed location and ith unknown facility location, and each $\bar{c}_{it}$ is a constant representing cost per unit distance between the ith and tth unknown facility location.

When the problem is reduced to that of finding the optimal location of a single facility on a two-dimensional plane, a mechanical model has been described by Burstall, et al. [3] and Haley [15]. The model has strings to each of which is attached a weight proportional to one of the c_i. A flat board is

*This research was supported by a grant from the Control Data Corporation. Use of the University of Wisconsin Computing Center was made possible through support, in part, from the National Science Foundation, other United States Government agencies and the Wisconsin Alumni Research Foundation through the University of Wisconsin Research Committee.

marked with a map showing the p fixed locations. At each fixed location is located a hole through which a string is passed. The strings are fastened together in a knot on the board. When the system reaches equilibrium. the knot represents the optimum location of the facility. The principal source of error in this method is friction, and the technique is limited to single facility location on a two-dimensional plane. A gradient-reducing numerical procedure has been described by Rogers and Vergin [25] based on a convexity proof by Haley [15]. Kuhn and Kuenne [19] have proposed an algorithm based on a perturbation technique and Bellman [2] has suggested a dynamic programming approach. Cooper [6] has developed heuristic solution methods for location-allocation problems. These appear to require large amounts of computation time. Other works are cited by Francis [13].

One application of facility location theory is that of multi-warehouse and factory location problems. This problem is usually treated as one in which a given set of possible locations are available and the main task is to choose a subset of these. A number of proposals have been made for solving the problem of multi-facility location by some form of linear programming [1]. Simulation methods have been employed [23]. Heuristic methods have been developed [18, 21], and a branch-bound algorithm is given by Efroymson and Ray [8]. A second type of facility location problem concerns the location of indivisible facilities [16], a common application of which is assigning work centers to location within an industrial plant.

In this article, we are concerned with the problem of locating facilities anywhere within the continuum of specified three-dimensional regions. Applications of the specific type of model with which we are concerned have been made. The model described by Burstall, et al. was applied to a warehouse location problem. In this case the c_{ij} values represent the total transportation cost per unit distance between the warehouse to be located and the jth source of destination of the goods. Cherniack and Schneider [5] have described a model of hospital location which seeks to minimize the total of all straight-line distances between the proposed hospital site and the centers of all postal zones in the areas to be served by the hospital. The straight-line distances are weighted by the density of patient populations in each zone. For computational purposes this model becomes a special case of the model we are considering here. The objective of the hospital location model is to minimize total patient travel distance outside the hospital. This same concept may be applied to other community service facilities such as post offices or fire stations.

Other applications [22, 24] exist in communication or electrical networks where the cost of the connecting lines between facilities is proportional to the total lengths of the lines that are used. The design of oil pipeline networks represent another application of this type of model.

Computational methods based on the assumption of costs proportional to straight-line distances which appear in the literature are generally restricted to the location of one facility in relation to a set of any number of other fixed locations. As far as this author can determine, successful models which represent situations in which there are several facilities to be located with costs present between the unlocated facilities have not been developed. Further, existing models other than a single facility model developed by the author [20] are concerned only with a situation where all facilities lie on a two-dimensional plane. Wiring and piping problems may be three-dimensional ones when applications are in fields such as electrical product wiring and petro-chemical plant piping layouts.

Another limitation of existing methods is the lack of consideration of the existence of spatial constraints. In many piping and wiring applications it may not be possible to have the connecting pipes or wires pass through certain regions due to existing or proposed physical barriers. In other applications it may be desirable to have some or all facilities lie within some specified region.

II. A CONVEX PROGRAMMING FORMULATION OF THE MULTI-FACILITY LOCATION PROBLEM

The computational method to be described here is more general than existing methods since it will optimally locate any number of facilities in three-dimensional space, it accounts for activity costs between all facilities, and will consider various types of spatial constraints.

Another serious problem both from a theoretical and practical computational viewpoint which has been ignored in the literature to date concerns the existence of the derivatives of the total cost function $C(x)$. In order to discuss this problem we first define some notation.

Consider two points, i and j, in three-space with coordinates (x_{1i}, x_{2i}, x_{3i}) and (x_{1j}, x_{2j}, x_{3j}), respectively. The distance between i and j is given by

$$f(x_{1i}, x_{2i}, x_{3i}, x_{1j}, x_{2j}, x_{3j}) = [(x_{1i}-x_{1j})^2+(x_{2i}-x_{2j})^2+(x_{3i}-x_{3j})^2]^{1/2}. \tag{2}$$

The following equation is identical to Eq. (2) except that the coordinates have been renumbered:

$$f(x) = [(x_1-x_2)^2+(x_3-x_4)^2+(x_5-x_6)^2]^{1/2}, \tag{3}$$

where $x=(x_1, \ldots, x_6)$ denotes a point in six-dimensional euclidean space, E^6. A direction in E^6 is given by the vector $u=(u_1, \ldots, u_6)$, where $|u|=1$. The first and second partial derivatives of f are given by D_i and D_{ij}, where

$$D_i=\frac{\partial f}{\partial x_i},\ D_{ij}=\frac{\partial^2 f}{\partial x_i \partial x_j};\ i, j=1, \ldots, 6.$$

The first directional derivative at a point x is given by $D_u f(x)$.

For the special case where one facility is to be located on a two-dimensional plane, Haley and Cooper have presented proofs of the convexity of f using differential conditions [7, 15]. A shortcoming of this approach is that f must have continuous partial derivatives of the second order [9]. The D_i of f are not continuous at any of the set of points $P=\{p|p_1=p_2,\ p_3=p_4,\ p_5=p_6\}$. To see this, consider the directional derivative $D_u f(x)$ at the point $p=(a, a, b, b, c, c)$. Clearly $p \epsilon P$. If λ is some number, then if $D_u f(p)$ existed it would be given by

$$\begin{aligned} D_u f(p) &= \lim_{\lambda\to 0}\frac{f(p+\lambda u)-f(p)}{\lambda} \\ &= \lim_{\lambda\to 0}\frac{|\lambda|}{\lambda}[(u_1-u_2)^2+(u_3-u_4)^2+(u_5-u_6)^2]^{1/2} \\ &= \begin{cases} [(u_1-u_2)^2+(u_3-u_4)^2+(u_5-u_6)^2]^{1/2}, & \lambda>0 \\ -[(u_1-u_2)^2+(u_3-u_4)^2+(u_5-u_6)^2]^{1/2}, & \lambda<0. \end{cases} \end{aligned}$$

This result shows there is no unique value for the partial derivatives at any point in the set P and hence they are non-existent at these points. In terms of locating points in E^3, points of discontinuity occur when any two facilities, or a facility and a fixed location occupy the same position.

A proof of the convexity of f which does not utilize the differential conditions is given by Kataoka [17]. Since the total cost function C given by Eq. (1) is a sum of convex functions of the same form as Eq. (3), C is convex. We may now state the problem as a convex programming problem.

Determine a vector x^* such that x^* maximizes $-C(x)$ subject to

$$g_1(x) \geqslant 0$$
$$\cdot\ \cdot\ \cdot\ \cdot\ \cdot$$
$$g_M(x) \geqslant 0$$
$$x \geqslant 0.$$

The $g_i(x)$, $i=1, \ldots, M$, are assumed to be concave functions since any set $K=\{x \| x \geqslant 0, g(x) \geqslant 0$, and $g(x)$ concave$\}$ is a convex set.

In general, due to the non-existence of the partial derivatives at any feasible point x such that $x \epsilon P$, it is not possible to guarantee the convergence to an optimal solution of any gradient reducing method such as those given in Refs. [6, 25]. During the course of numerical computations it has been found that as two points approach each other, the derivatives may quickly change from large positive values to large negative values. Rogers and Vergin [25] have called these derivatives "knife edged." The result is an oscillation effect producing very slow convergence of the solution. If the final optimal solution requires that two points are to coincide, any type of gradient method cannot converge. The Method of Fitted Functions developed in section IV of this paper solves this discontinuity problem, is computationally efficient, and guarantees that a unique optimal solution will be reached within any degree of accuracy in all cases.

III. PERMISSIBLE CONSTRAINT EQUATIONS

In order to insure convergence of solution methods using gradient techniques, the constraint equations must define a convex set. Linear constraint equations, of course, satisfy this requirement, making it possible to constrain the facilities to lie within certain boundaries. For example, if a facility is being located within a rectangular volume and it is to be confined to the upper right rear corner, then constraint equations of the form

$$\begin{aligned} x_1 - D_1 &\geqslant 0 \\ x_2 - D_2 &\geqslant 0 \\ x_3 - D_3 &\geqslant 0 \end{aligned} \tag{5}$$

or

$$a_1x_1 + a_2x_2 + a_3x_3 - D_4 \geqslant 0 \tag{6}$$

can be employed. (See Figures 1(a) and (b). In this case the constants D_i, $i=1, 2, 3, 4$, and a_i, $i=1, 2, 3$, are positive.)

A second type of constraint which is permissible is that which constrains a facility to lie within a certain spherical region. For example, if it was desired to keep it within a distance r of a fixed point x', the constraint

$$[(x_1 - x_1')^2 + (x_2 - x_2')^2 + (x_3 - x_3')^2]^{1/2} \leqslant r$$

would be required. This can be written as

$$r - [(x_1 - x_1')^2 + (x_2 - x_2')^2 + (x_3 - x_3')^2]^{1/2} \geqslant 0. \tag{7}$$

Figure 2 illustrates a constraint of this type.

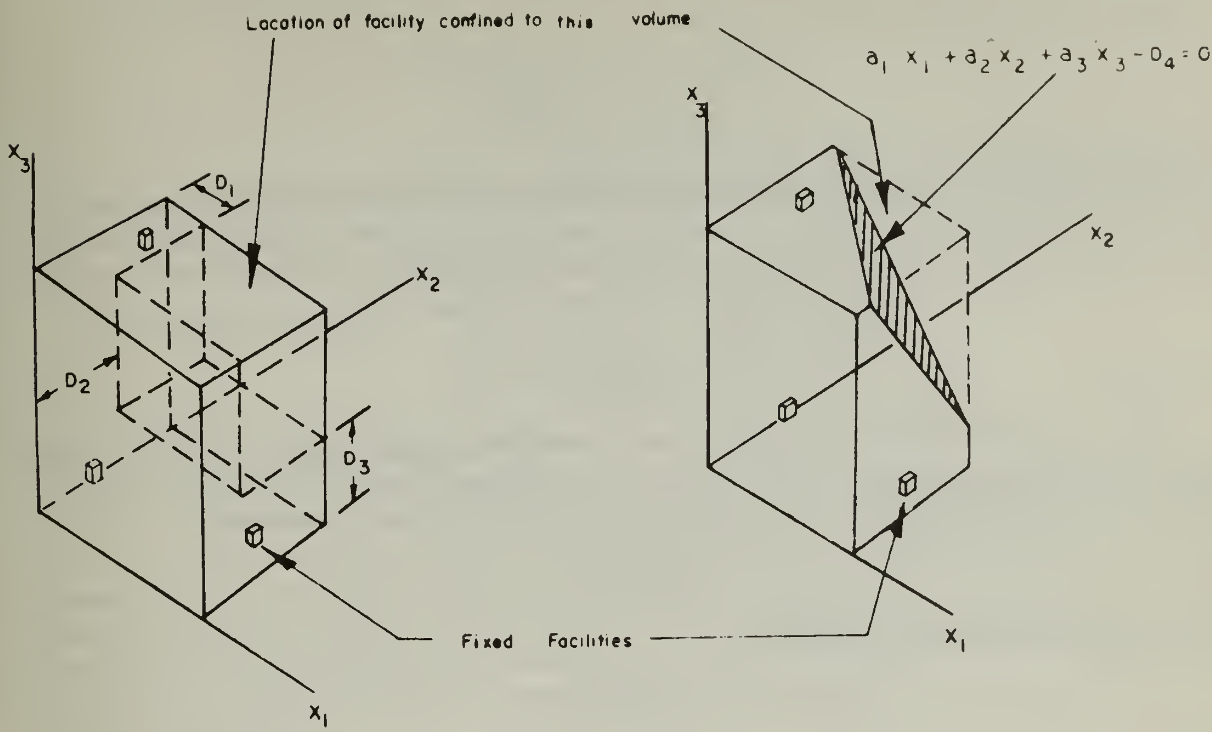


Figure 1(a)

Figure 1(b)

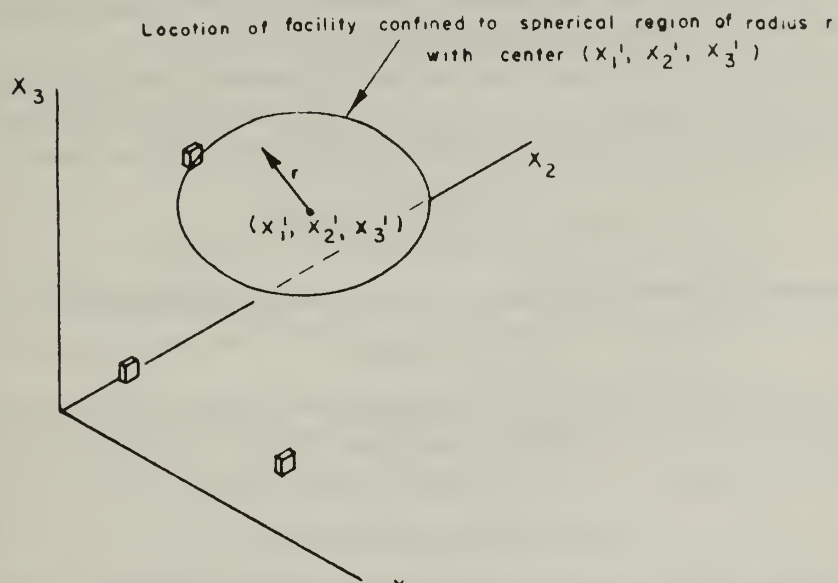


Figure 2

Since the negative of any convex function is concave, the function on the left side of Eq. (7) is concave.

IV. COMPUTATION: THE METHOD OF FITTED FUNCTIONS

The computational technique employed was an extended version of Carroll's created response-surface technique [4]. Convergence proofs and extensions of this method are given by Fiacco and McCormick [10–12].

The created response surface is given by

$$p(x, R) = C(x) + \sum_{i=1}^{M} \frac{R}{g_i(x)},$$

where $R \geqslant 0$ and the $g_i(x)$ are the constraint functions, $i = 1, \ldots, M$.

The first step of Carroll's method is to choose a value of R and an initial feasible point. Next, the gradient is calculated at this point and a search conducted along the gradient until an improved solution is found. Holding R constant, the process is repeated as many times as necessary until the criterion

$$\left| \frac{\partial P}{\partial x_{ij}} \right| \leqslant \delta > 0, \qquad i = 1, 2, 3; j = 1, \ldots, m$$

is satisfied.

Next, a new smaller value of R is chosen and the gradient reducing process is repeated. The final solution from the preceding value of R is used as the initial feasible point for the new R value. A sequence of values for R, R_1, R_2, . . ., is chosen so that the sequence meets the criterion

$$\lim_{i \to \infty} R_i = 0$$

In our numerical examples, the sequence

$$R_i = \frac{R_{i-1}}{4^i},$$

$R_1 = 0.5$, and $\delta = 0.01$ gave good results.

The method of choosing successive values of x in the gradient reducing procedure for each value of R is based on a method proposed by Goldstein [14].

Let the value of the created response surface for any values of R and x be given by $P(x^k, R_i)$ where R_i is the ith value of R, and x^k is the value of x selected at the kth step in the gradient reduction process when $R = R_i$. Let $\nabla_x P(x^k, R_i)$ be the gradient vector of $P(x^k, R_i)$ and define

$$w^k = \frac{-P(x^k, R_i) \nabla P(x^k, R_i)}{|\nabla P(x^k, R_i)|^2}. \tag{8}$$

Given a value of R, successive values for x were determined from the equation

$$x^{k+1} = x^k + d_k w^k, \tag{9}$$

where d_k is chosen so that $P(x^{k+1}, R_i) < P(x^k, R_i)$. Initially d_0 is set equal to a constant, say 1. The iterations are continued with $d_{k+1} = d_k$ unless at some value of k, $P(x^{k+1}, R_i) \geqslant P(x^k, R_i)$. In this case, d_{k+1} is chosen from the equation

$$d_{k+1} = d_k / 2^n, n = 1, 2, \ldots, \tag{10}$$

where successively larger values of n are chosen, starting with $n = 1$, until the criterion

$$P(x^{k+1}, R_i) < (x^k, R_i)$$

is met. The sequence $P(x^{k+1}, R_i)$, $k = 0, 1, \ldots$, is monotonic decreasing. A convergence proof for the method is given by Goldstein.

The computation of successive values of R is terminated when the value of $C(x)$ comes within a predetermined interval of the exact optimal solution. The use of the dual problem to compute this interval is described by Fiacco and McCormick [10].

The dual problem is given as:

$$\text{maximize } G(x, \mu) = C(x) - \sum_{i=1}^{n} \mu_i g_i(x)$$

$$\text{subject to } \nabla_x G(x, \mu) = \vec{0}, \ \mu_i \geq 0.$$

If $x(R_k)$ is the point for which $P(x, R_k)$ is minimized, then the dual solution variables are given by

$$\mu_i(R_k) = R_k / g_i^2[x(R_k)]; \ i = 1, \ldots, n.$$

The computation process is terminated when for some j, $j = 1, 2, \ldots,$

$$\frac{C[x(R_j)] - G[x(R_j), \mu(R_j)]}{C[x(R_j)]} \leq \epsilon.$$

where $\epsilon > 0$ is some predetermined number.

The condition that $\nabla_x P[x(R_k), R_k] = \vec{0}$ for each R_k required by the Carroll-Fiacco-McCormick solution procedure cannot be met in general due to the nature of the derivatives of $C(x)$. The method developed to circumvent this computational difficulty approximates $C(x)$ with a sequence of functions which is uniformly convergent to $C(x)$; each function in the sequence has continuous first partial derivatives everywhere. The optimal solution to the problem is computed for each term in the sequence. In actual computations the sequence is terminated when it approximates $C(x)$ with some predetermined degree of accuracy.

The Fitted Function method modifies the $C(x)$ function over certain parts of the domain of x by adding the following rule to the computational process. Given the ith value of R given by R_i, x^k, and $\nabla_x P(x^k, R_i)$, each time x^{k+1} has been determined, the following tests are made for some $T > 0$ and $i = 1, \ldots, m$; $j = 1, \ldots, p$; $t = 1, \ldots, m$:

$$\text{Is } C_{it}^t = [(x_{1i} - x_{1t})^2 + (x_{2i} - x_{2t})^2 + (x_{3i} - x_{3t})^2]^{1/2} \geq T, \tag{11}$$

and

$$C_{ij}^j = [(a_{1j} - x_{1i})^2 + (a_{2j} - x_{2i})^2 + (a_{3j} - x_{3i})^2]^{1/2} \geq T? \tag{12}$$

If the answers to these tests are all yes, the computational process is resumed as before. If, however, for some i and t values, say i' and t', $C_{i',t'}^t < T$, then $C_{i',t'}^t$ is replaced in $C(x)$ by $\bar{C}_{i',t'}^t$ where $\bar{C}_{i',t'}^t$ is given by

$$\bar{C}_{i',t'}^t = (1/2T)[(x_{1i'} - x_{1t'})^2 + (x_{2i'} - x_{2t'})^2 + (x_{3i'} - x_{3t'})^2] + T/2. \tag{13}$$

Similarly, if for some i and j values, say i' and j',

$$C_{i',j'}^j < T, \text{ then } C_{i',j'}^j \text{ is replaced in } C(x) \text{ by } \bar{C}_{i',j'}^j, \text{ where}$$

$$\bar{C}_{i',j'}^j = (1/2T)[(a_{1j} - x_{1i})^2 + (a_{2j} - x_{2i})^2 + (a_{3j} - x_{3i})^2] + T/2. \tag{14}$$

If, at any later time during the computations, the tests given by Eqs. (11) and (12) are satisfied for these same i, t, or j values, the $\bar{C}_{i,t}^t$ or $\bar{C}_{i,j}^j$ functions are replaced by the $C_{i,t}^t$ or $C_{i,j}^j$ functions which they originally displaced. Use of the $\bar{C}_{i,t}^t$ and $\bar{C}_{i,j}^j$ substitutions creates a revised total cost or primal objective function. We will refer to this new function as the fitted function to $C(x)$, denoted by $C(x; T)$. The created response function corresponding to $C(x; T)$ is given by $P(x, R, T)$.

The computational procedure incorporates a sequence of T values. An initial value of T is chosen. The Carroll-Fiacco-McCormick optimization procedure is performed on $C(x;\,T)$ with a given sequence of R values. For each R value a stopping criterion calculation is made (formulation of the stopping criterion is given later in the paper). If, at the end of the R sequence the stopping criterion has not been met, a new smaller value of T is chosen and the optimization calculations are repeated. Successive values of T are chosen and the optimization calculations are repeated until the stopping criterion is met.

We can justify the creation of $C(x;\,T)$ by proving the following:

1. $C(x;\,T)$ is continuous everywhere.

2. $C(x;\,T)$ has continuous first partial derivatives everywhere.

3. When a sequence of values for T, $\{T_1, T_2, \ldots\}$, is chosen which meets the criterion $\lim_{i\to\infty} T_i = 0$, the sequence of functions $\{C(x;\,T_i)\}$ converges uniformly to $C(x)$.

Properties 1, 2 are necessary in order to ensure that the optimization computation procedure will minimize the $P(x,\,R,\,T)$ function. The third property is used in the formulation of the stopping criterion which guarantees that the final solution will be within some predetermined interval of the global optimum solution of the original primal problem. Before discussing the stopping criterion, we first prove the three properties of $C(x;\,T)$ given above.

1. Proof That $C(x;\,T)$ Is Everywhere Continuous:

Since the $\bar{C}_{it}^t$ and $\bar{C}_{ij}^t$ functions are polynomials they are everywhere continuous. Thus, $C(x;\,T)$ does not have the points of discontinuity that characterizes $C(x)$. If we consider the function C_{it}^t and its "replacement," $\bar{C}_{it}^t$, a possible set of points of discontinuity occur on the set $S=\{x;\,C_{i,t}^t=\bar{C}_{i,t}^t=T\}$; however, since both C_{it}^t and $\bar{C}_{it}^t$ are equal on S and continuous everywhere defined by $C(x;\,T)$, $C(x;\,T)$ is continuous at S. A similar argument holds for the function C_{ij}^j and its "replacement" $\bar{C}_{ij}^j$.

2. Proof That $C(x;T)$ Has Continuous First Partial Derivatives Everywhere:

Consider the function C_{it}^t and its replacement $\bar{C}_{it}^t$ when $x\epsilon S$. For any k, $k=1, 2, 3$,

$$
\begin{aligned}
\partial C_{it}^t/\partial x_{ki} &= \left[\sum_{n=1}^{3}(x_{ni}-x_{nt})^2\right]^{-1/2}(x_{ki}-x_{kt}) \\
&= T^{-1}(x_{ki}-x_{kt}). \qquad (15)\\
&= \partial \bar{C}_{it}^t/\partial x_{ki}.
\end{aligned}
$$

Similarly

$$
\begin{aligned}
\partial C_{it}^t/\partial x_{kt} &= -\left[\sum_{n=1}^{3}(x_{ni}-x_{nt})^2\right]^{-1/2}(x_{ki}-x_{kt}) \\
&= -T^{-1}(x_{ki}-x_{kt}) \qquad (16)\\
&= \partial \bar{C}_{it}^t/\partial x_{kt}.
\end{aligned}
$$

In the same way, it can be proved that the partial derivatives with respect to any variable in any pair of C_{ij}^t and $\bar{C}_{ij}^t$ functions are equal on S. All of these partial derivative functions are continuous everywhere defined by $C(x;\,T)$ and therefore are continuous on S.

3. Proof That $C(x;\,T)$ Is Uniformly Convergent to $C(x)$:

Define

$$f_1(x) = 1/2T[(x_1-x_2)^2+(x_3-x_4)^2+(x_5-x_6)^2]+T/2. \qquad (17)$$

Define the function $S=f(y)$ for any y such that $S\epsilon[0, T]$. f is given by Eq. 3.
For any y in the domain of the function $S=f(y)$, it follows that

(18) $$f_1(y)=S^2/2T+T/2.$$

Define $g(S)$, the difference between the actual and fitted functions, as

(19) $$g(S)=g_1(y)=f_1(y)-f(y)=\frac{S^2}{2T}+\frac{T}{2}-S.$$

Then

(20) $$g'(S)=\frac{S}{T}-1, \text{ and}$$

(21) $$g''(S)=T^{-1}.$$

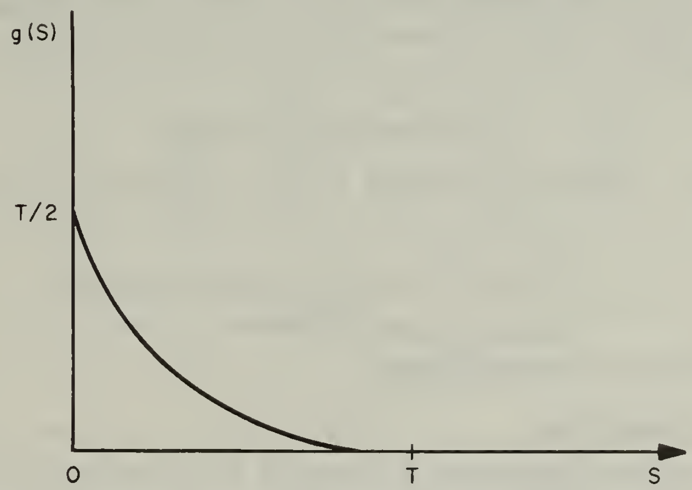


Figure 3

Since $g'(0)=-1$, $g'(S)<0$ for $0<S<T$, $g'(T)=0$, and $g''(S)>0$ for $0\leqslant S\leqslant T$, it follows that $g(S)$ is a strictly decreasing convex function in the interval $[0, T]$, as shown in Figure 3.

Let $c_m=\max\limits_{i,j} c_{ij}$. Then

(22) $$|C(x)-C(x;T)|\leqslant\left(\sum_{i=1}^{m}\sum_{t=1}^{m}\bar{c}_{it}+2mc_m\right)T/2=AT,$$

where $$A=\left(\frac{1}{2}\right)\sum_{i=1}^{m}\sum_{t=1}^{m}\bar{c}_{it}+2mc_m.$$

Since $\lim\limits_{i\to\infty} T_i=0$, for each $\epsilon>0$, there exists an N such that for $i>N$ $|T_i|<\epsilon/A$ since $A>0$.

Therefore, for all $i>N$,

(23) $$|C(x)-C(x;T)|=AT<\epsilon \text{ for all real } x.$$

This completes the proof that $C(x;T)$ converges uniformly to $C(x)$.

V. STOPPING CRITERION.

The computation of successive values of R and T is terminated when the value of $C(x;T)$ comes within a predetermined interval of the exact optimal solution. The dual problem is used to compute

this interval. It is given by

$$\text{(24)} \qquad \text{maximize } G(x, \mu T) = C(x; T) - \sum_{h=1}^{k} \mu_h g_h(x)$$

$$\text{subject to } \nabla_x G(x, \mu) = \vec{0},$$

$\mu \geqslant \overline{0}$, where $\mu = (\mu_1, \ldots, \mu_k)$ is the vector of dual variables.

When $x(R_i, T_j)$ is defined as the point at which $P(x, R_i, T_j)$ is minimized, Fiacco and McCormick prove that $\mu_j(R_i) = R_i / g_j^2[x(R_i)]$, and

$$\text{(25)} \qquad G[x(R_i), \mu(R_i), T_j] = C[x(R_i);+ T_j] - R_i \sum_{h=0}^{M} 1/g_h[x(R_i)],$$

where $G[x(R_i), \mu(R_i), T_j]$ is the value of the dual problem when $P(x, R_i, T_j)$ is minimized.

For any $x(R_i, T_j)$, the final optimal value of $C(x; T_j)$ lies in the interval bounded above by $C[x(R_i); T_j]$ and below by $G[x(R_i), \mu(R_i), T_j]$. Since $|C(x) - C(x; T_j)| \leqslant AT_j$, a lower bound on $C(x)$ given $x(R_i)$ and T_j is given by

$$G[x(R_i), \mu(R_i), T_j] - AT_j.$$

In our numerical examples, computations were terminated when the following criterion was satisfied:

$$\text{(24)} \qquad S.C. = \frac{C[x(R_i)] - \{[Gx(R_i), \mu(R_i), T_j] - AT_j\}}{C[x(R_i)]} \leqslant \gamma,$$

for some $\gamma > 0$, where $S.C. \equiv$ stopping criterion.

VI. NUMERICAL RESULTS*

The algorithm has been programmed in Fortran IV to accommodate any number of fixed and variable facilities for two and three dimensional unconstrained problems. In general, each constrained problem is a special case requiring some additional individual programming. About 20 different problems have been run and the computational pattern is similar in each case. The algorithm converges rapidly to a small neighborhood of the optimal solution. Tables 1 to 4 give the system parameters and computational results for a constrained problem with 10 fixed facility locations and 5 facilities to be located. The single constraint was $x_{11} + x_{21} \leqslant 24$. In less than 19 seconds the algorithm converges to a solution which is within 0.624 percent of the optimal solution. A three-dimensional problem corresponding to this example with the sāme number of fixed and variable facilities requires approximately 10 percent more computing time. A two-dimensional unconstrained problem with 50 fixed facilities and 10 variable facilities required 38 seconds to come within 1 percent of the optimum solution and 90 seconds to come within .1 percent of the optimum.

TABLE 1

Location of Ten Fixed Facilities

j	1	2	3	4	5	6	7	8	9	10
(x_{1j}, x_{2j})	(2, 5)	(10, 20)	(10, 10)	(20, 20)	(15, 15)	(21, 1)	(21, 11)	(1, 31)	(10, 17)	(11, 17)

*All numerical examples were run on the CDC 3600 computer at the University of Wisconsin Computing Center.

TABLE 2. *Cost Per Unit Distance between the jth Fixed Location and the ith Unknown Facility Location,* c_{ij}

i \ j	1	2	3	4	5	6	7	8	9	10
1	0.16	0.56	0.16	0.16	0.14	0.58	0.18	0.18	0.10	0.18
2	0.70	0.60	0.20	0.90	0.70	0.00	0.80	0.60	0.10	0.40
3	0.00	0.50	0.10	0.60	0.20	0.30	0.50	0.40	0.50	0.90
4	0.09	0.03	0.04	0.04	0.05	0.02	0.08	0.35	0.06	0.09
5	0.30	0.50	0.50	0.30	1.00	0.10	0.30	0.50	0.10	0.50

TABLE 3. *Cost Per Unit Distance between the ith and tth Unknown Facility Location,* c_{it}

i \ j	1	2	3	4	5
1	....	0.10	0.00	0.00	0.00
2	0.10	....	0.00	0.00	0.10
3	0.00	0.10	....	0.00	0.00
4	0.10	0.10	0.00	....	0.10
5	0.00	0.10	0.00	0.00	....

TABLE 4. *Computation Results for Test Problem*

$T=1.0$					
i		1	2	3	4
R_i	0.00	0.25	0.0625	0.0156	0.0039
x_{11}	0.00	9.08	9.20	9.29	9.33
x_{21}	0.00	14.05	14.34	14.48	14.56
x_{12}	0.00	9.61	9.65	9.66	9.67
x_{22}	0.00	14.02	14.13	14.17	14.20
x_{13}	0.00	9.09	9.09	9.09	9.09
x_{23}	0.00	20.22	20.22	20.22	20.22
x_{14}	0.00	10.13	10.17	10.18	10.19
x_{24}	0.00	14.38	14.54	14.57	14.60
x_{15}	0.00	10.54	10.54	10.54	10.54
x_{25}	0.00	16.47	16.47	16.47	16.47
$\partial P/\partial x_{11}$		0.0014	−0.0089	0.0045	0.0077
$\partial P/\partial x_{21}$		−0.0019	−0.0098	−0.0024	0.0016
$\partial P/\partial x_{12}$		−0.0002	0.0004	0.0002	−0.0006
$\partial P/\partial x_{22}$		−0.0031	−0.0007	−0.0030	−0.0033
$\partial P/\partial x_{13}$		−0.0000	−0.0000	0.0000	−0.0000
$\partial P/\partial x_{23}$		−0.0000	−0.0000	0.0000	0.0000
$\partial P/\partial x_{14}$		−0.0018	−0.0004	−0.0018	−0.0015
$\partial P/\partial x_{24}$		−0.0092	−0.0033	−0.0063	−0.0061
$\partial P/\partial x_{15}$		−0.0063	−0.0000	0.0001	0.0000
$\partial P/\partial x_{25}$		−0.0005	0.0000	−0.0001	−0.0000
$C(x)$	499.169	251.092	250.959	250.891	250.859
$C[x(R_i);T]$		251.165	251.033	250.966	250.933
$G[x(R_i),\mu(R_i),T]$		250.876	250.897	250.899	250.900
S.C.		0.00624	0.00583	0.00555	0.00542
Elapsed Computing Time (Sec.)		18.564	31.470	36.717	51.631

TABLE 4—Continued

$T=0.5$					
i		1	2	3	4
R_i		0.25	0.0625	0.0156	0.0039
x_{11}	9.33	9.15	9.32	9.40	9.44
x_{21}	14.56	14.01	14.23	14.37	14.44
x_{12}	9.67	9.58	9.62	9.65	9.66
x_{22}	14.20	14.05	14.12	14.19	14.23
x_{13}	9.09	9.24	9.24	9.24	9.24
x_{23}	20.22	20.16	20.16	20.16	20.16
x_{14}	10.19	9.93	9.94	9.96	9.97
x_{24}	14.60	14.32	14.33	14.40	14.43
x_{15}	10.54	10.42	10.42	10.43	10.43
x_{25}	16.47	16.69	16.69	16.69	16.69
$\partial P/\partial x_{11}$	17.7192	−0.0013	0.0088	0.0024	0.0096
$\partial P/\partial x_{21}$	17.8003	0.0022	0.0004	−0.0045	0.0017
$\partial P/\partial x_{12}$	0.0139	0.0014	−0.0013	−0.0005	−0.0013
$\partial P/\partial x_{22}$	−0.0674	0.0043	−0.0057	−0.0048	−0.0066
$\partial P/\partial x_{13}$	−0.0552	−0.0001	−0.0002	−0.0001	−0.0002
$\partial P/\partial x_{23}$	0.0135	0.0001	0.0001	0.0001	0.0001
$\partial P/\partial x_{14}$	0.0285	0.0049	−0.0021	−0.0023	−0.0028
$\partial P/\partial x_{24}$	0.0153	0.0093	−0.0069	−0.0091	−0.0100
$\partial P/\partial x_{15}$	0.0733	0.0001	0.0000	0.0000	−0.0000
$\partial P/\partial x_{25}$	−0.2799	−0.0001	−0.0000	−0.0000	0.0000
$C(x)$	250.859	251.032	250.899	250.834	250.801
$C[x(R_i); T]$	250.859	251.033	250.904	250.840	250.807
$G[x(R_i), \mu(R_i), T]$		250.735	250.765	250.773	250.774
S.C.		0.00148	0.00083	0.00053	0.00041
Elapsed Computing Time (Sec.)	51.938	60.571	63.420	68.485	75.176

$T=0.25$					
i		1	2	3	4
R_i		0.25	0.0625	0.0156	0.0039
x_{11}	9.44	9.20	9.41	9.48	9.52
x_{21}	14.44	13.97	14.15	14.28	14.36
x_{12}	9.66	9.56	9.61	9.64	9.66
x_{22}	14.23	14.02	14.10	14.19	14.24
x_{13}	9.24	9.24	9.24	9.24	9.24
x_{23}	20.16	20.16	20.16	20.16	20.16
x_{14}	9.97	9.84	9.81	9.82	9.84
x_{24}	14.43	14.21	14.21	14.30	14.35
x_{15}	10.43	10.42	10.42	10.42	10.42
x_{25}	16.69	16.69	16.69	16.69	16.69
$\partial P/\partial x_{11}$	17.4006	−0.0018	0.0007	0.0026	0.0039
$\partial P/\partial x_{21}$	17.4572	0.0019	−0.0026	−0.0048	−0.0032
$\partial P/\partial x_{12}$	0.0090	0.0001	−0.0009	−0.0001	−0.0005
$\partial P/\partial x_{22}$	−0.0534	0.0019	−0.0077	−0.0058	−0.0059
$\partial P/\partial x_{13}$	−0.0002	0.0000	−0.0004	−0.0002	−0.0002
$\partial P/\partial x_{23}$	0.0001	−0.0000	0.0001	0.0001	0.0001
$\partial P/\partial x_{14}$	0.0198	0.0093	0.0053	−0.0017	−0.0024
$\partial P/\partial x_{24}$	0.0051	0.0100	−0.0070	−0.0094	−0.0095
$\partial P/\partial x_{15}$	−0.0000	0.0002	0.0001	0.0001	0.0001
$\partial P/\partial x_{25}$	0.0000	−0.0001	−0.0000	−0.0001	−0.0000
$C(x)$	250.801	251.026	250.893	250.827	250.794
$C[x(R_i); T]$	250.801	251.026	250.894	250.829	250.796
$G[x(R_i), \mu(R_i), T]$		250.725	250.754	250.812	250.763
S.C.		0.00120	0.00070	0.00021	0.00027
Elapsed Computing Time (Sec.)	75.487	87.767	90.910	96.937	103.016

TABLE 4—Continued

$T = 0.1250$					
i		1	2	3	4
R_i		0.25	0.0625	0.0156	0.0039
x_{11}	9.52	9.19	9.43	9.54	9.58
x_{21}	14.36	13.98	14.13	14.23	14.30
x_{12}	9.66	9.57	9.59	9.64	9.66
x_{22}	14.24	14.03	14.11	14.18	14.24
x_{13}	9.24	9.24	9.24	9.24	9.24
x_{23}	20.16	20.16	20.16	20.16	20.16
x_{14}	9.84	9.81	9.77	9.74	9.76
x_{24}	14.35	14.20	14.20	14.23	14.30
x_{15}	10.42	10.42	10.42	10.42	10.42
x_{25}	16.69	16.69	16.69	16.69	16.69
$\partial P/\partial x_{11}$	0.0039	−0.0031	−0.0047	0.0031	0.0071
$\partial P/\partial x_{21}$	−0.0032	0.0033	0.0030	−0.0042	−0.0014
$\partial P/\partial x_{12}$	−0.0005	0.0006	−0.0008	0.0015	0.0000
$\partial P/\partial x_{22}$	−0.0059	0.0036	−0.0025	−0.0066	−0.0069
$\partial P/\partial x_{13}$	−0.0002	0.0002	−0.0002	−0.0004	−0.0002
$\partial P/\partial x_{23}$	0.0001	−0.0001	0.0001	0.0002	0.0001
$\partial P/\partial x_{14}$	−0.0024	0.0066	0.0097	0.0030	−0.0009
$\partial P/\partial x_{24}$	−0.0095	0.0099	−0.0018	−0.0098	−0.0090
$\partial P/\partial x_{15}$	0.0001	0.0001	0.0003	0.0003	0.0002
$\partial P/\partial x_{25}$	−0.0000	−0.0000	−0.0001	−0.0001	−0.0001
$C(x)$	250.794	251.025	250.890	250.825	250.792
$C[x(R_i); T]$	250.796	251.025	250.890	250.825	250.792
$G[x(R_i), \mu(R_i), T]$		250.724	250.747	250.758	250.759
S.C.		0.00120	0.00057	0.00034	0.00020
Elapsed Computing Time (Sec.)	103.016	115.331	120.369	124.493	131.032

ACKNOWLEDGMENTS

The author would like to thank George Wesolowsky, Ph. D. student in quantitative analysis at the University of Wisconsin, for writing the Fortran programs and supervising all phases of the computing process that made it possible to do the numerical examples used in this research. I would also like to thank the reviewer for pointing out the errors and omissions in the original manuscript.

REFERENCES

[1] Baumol, W. J. and P. Wolfe, "A Warehouse Location Problem," Operations Research, *6*, 252–263 (1958).

[2] Bellman, R., "An Application of Dynamic Programming to Location—Allocation Problems," SIAM Review, *7*, 126–128 (1965).

[3] Burstall, R. A., R. A. Leaver, and J. E. Sussans, "Evaluation of Transport Cost For Alternative Factory Sites," Operational Research Quarterly, *13*, 345–354 (1962).

[4] Carroll, C. W., "The Created Response Surface Technique for Optimizing Non-Linear Restrained Systems," Operations Research, *9*, 169–184 (1961).

[5] Cherniack, H. D. and J. B. Schneider, "A New Approach to the Delineation of Hospital Service Areas," Discussion Paper Series No. 16, Regional Science Research Institute, Seattle, Washington.

[6] Cooper, L., "Heuristic Methods for Location—Allocation Problems," SIAM Review, *6*, 37–53 (1964).

[7] Cooper, L., "Location—Allocation Problems," Operations Research, *11*, 331–343 (1963).

[8] Efroymson, M. A. and T. L. Ray, "A Branch-Bound Algorithm for Plant Location," Operations Research, *14*, 213–235 (1966).

[9] Eggleston, H. G., *Convexity* (Cambridge University Press, Cambridge, 1963), p. 51.

[10] Fiacco, A. V. and G. P. McCormick, "The Sequential Unconstrained Minimization Technique for Nonlinear Programming, A Primal-Dual Method," Management Science, *10*, 360–364 (1964).

[11] ______, "Computational Algorithm for the Sequential Unconstrained Minimization Technique for Nonlinear Programming," Management Science, *10*, 601–617 (1964).
[12] ______, "Extension of SUMT for Nonlinear Programming: Equality Constraints and Extrapolation," Management Science, *12*, 816–828 (1966).
[13] Francis, R. L., "Conditions for Optimum-Property Facility Design," Operations Research, *15*, 448–466 (1967).
[14] Goldstein, A. A., "Cauchy's Method of Minimization," Numerische Mathematik, *4*, 146–150 (1962).
[15] Haley, K. B., "The Siting of Depots," The International Journal of Production Research, *2*, 41–45 (1963).
[16] Hillier, F. S. and M. M. Connors, "Quadratic Assignment Problem Algorithms and the Location of Indivisible Facilities," Management Science, *13*, 42–57 (1966).
[17] Kataoka, S., "A Stochastic Programming Model," Econometrica, *31*, 181–196 (1963).
[18] Kuehn, A. A. and M. J. Hamburger, "A Heuristic Program for Locating Warehouses," Management Science, *9*, 643–666 (1963).
[19] Kuenne, R. E. and H. W. Kuhn, "An Efficient Algorithm for the Numerical Solution of the Generalized Weber Problem in Spatial Economics," Journal of Regional Science, *4*, 21–33 (1962).
[20] Love, R. F., "A Nonlinear Programming Approach to Facilities Location Problems," Journal of the Canadian Operational Research Society, *5*, 136–143 (1967).
[21] Manne, A. S., "Plant Location Under Economics-of-Scale: Decentralization and Computation," Management Science, *11*, 213–235 (1964).
[22] Palermo, F. P., "A Network Minimization Problem," IBM Journal (Oct. 1961).
[23] Shycon, H. N. and R. B. Maffei, "Simulation—Tool for Distribution," Harvard Business Review, *41*, 65–75 (1960).
[24] Steinberg, L., "The Blackboard Wiring Problem: A Placement Algorithm," SIAM Review, *3*, 37–50 (1960).
[25] Vergin, R. C. and J. D. Rogers, "An Algorithm and Computational Procedure for Locating Economic Facilities," Management Science, *13*, 240–254 (1967).

* * *

A DECOMPOSABLE TRANSSHIPMENT ALGORITHM FOR A MULTIPERIOD TRANSPORTATION PROBLEM*

M. Bellmore, W. D. Eklof,†and G. L. Nemhauser

The Johns Hopkins University

ABSTRACT

A dynamic version of the transportation (Hitchcock) problem occurs when there are demands at each of n sinks for T periods which can be fulfilled by shipments from m sources. A requirement in period t_2 can be satisfied by a shipment in the same period (a linear shipping cost is incurred) or by a shipment in period $t_1 < t_2$ (in addition to the linear shipping cost a linear inventory cost is incurred for every period in which the commodity is stored). A well known method for solving this problem is to transform it into an equivalent single period transportation problem with mT sources and nT sinks.

Our approach treats the model as a transshipment problem consisting of T, m source$-n$ sink transportation problems linked together by inventory variables.

Storage requirements are proportional to T^2 for the single period equivalent transportation algorithm, proportional to T, for our algorithm without decomposition, and independent of T for our algorithm with decomposition. This storage saving feature enables much larger problems to be solved than were previously possible. Futhermore, we can easily incorporate upper bounds on inventories. This is not possible in the single period transportation equivalent.

INTRODUCTION

The classical transportation problem can be stated as:

PROBLEM 1: Minimize $$\sum_{i=1}^{m}\sum_{j=1}^{n} C_{ij}X_{ij}$$

subject to

$$\sum_{j=1}^{n} X_{ij} \leq a_i \qquad i=1, 2, \ldots, m$$

$$\sum_{i=1}^{m} X_{ij} \geq b_j \qquad j=1, 2, \ldots, n$$

$$X_{ij} \geq 0 \qquad \text{all } i \text{ and } j,$$

where

$X_{ij}=$ the amount shipped from source i to sink j,

$a_i=$ the amount available at source i,

$b_j=$ the amount demanded at sink j, and

$C_{ij}=$ the linear cost applicable to X_{ij}.

The classical problem deals with shipping plans only in a single period. If the availabilities and demands are seasonal or the costs vary from one period to the next, and it is feasible to carry inventories

*Partially supported by National Bureau of Standards Contract CST−1279.

†Presently with Esso Mathematics & Systems Inc., Florham Park, New Jersey.

at the sources and/or sinks then the classical model is inadequate to describe the feasible shipping plans over a series of periods.

For example, if the shipping costs increase with time and are larger than the inventory costs or, if availabilities are smaller than demands in later periods, it will be desirable to maintain inventory at the sinks. A multiperiod formulation of the transportation problem with sink inventories permitted can be stated as:

PROBLEM 2: Minimize
$$\sum_{t=1}^{T}\sum_{i=1}^{m}\sum_{j=1}^{n} C_{ij}^{t} X_{ij}^{t} + \sum_{t=2}^{T}\sum_{j=1}^{n} d_j^t I_j^t$$

subject to
$$\sum_{j=1}^{n} X_{ij}^{t} \leq a_i^t \qquad \text{all } i \text{ and } t$$

$$\sum_{i=1}^{m} X_{ij}^{t} + I_j^t - I_j^{t+1} \geq b_j^t \quad \text{all } j \text{ and } t$$

$$0 \leq I_j^t \leq N_j^t \qquad \text{all } j \text{ and } t$$

$$X_{ij}^{t} \geq 0 \qquad \text{all } i, j \text{ and } t,$$

where X_{ij}^t, C_{ij}^t, a_i^t, and b_j^t are defined as above except that each now carries a superscript to denote the period to which they refer:

I_j^t = the amount of inventory carried from period $t-1$ to period t at sink j.*

d_j^t = the linear inventory cost applicable to I_j^t.

N_j^t = upper bound on I_j^t.

Bowman [1] and Kriebel [7] give a solution technique for Problem 2 when $N_j^t = \infty$ for all j, and t.† It is based on creating a large single period problem with mT sources and nT sinks. By defining

$$\hat{a}_{pm+i} = a_i^{p+1} \quad i = 1, \ldots, m, \; p = 0, 1, \ldots, T-1,$$

$$\hat{b}_{qn+j} = b_j^{q+1} \quad j = 1, \ldots, n, \; q = 0, 1, \ldots, T-1,$$

and

$$\hat{C}_{pm+i,\, qn+j} = \begin{cases} C_{ij}^{p+1} & p = q \\ C_{ij}^{p+1} + \sum_{\tau=p+2}^{q+1} d_j^\tau & q > p, \\ \infty & q < p \end{cases}$$

Problem 2 can be written in the format of Problem 1.

In this format

$$I_j^t = \sum_{q=t-1}^{T-1} \sum_{k=1}^{(t-1)m} \hat{X}_{k,\, qn+j},$$

where $\hat{X}_{pm+i,\, qn+j}$ is a shipment from source i in period $p+1$ designated for use by sink j in period

---

*It is assumed without loss of generality that $I_j^1 = 0$ for all j.

† Actually Bowman's model is in the context of a production smoothing problem which he shows to be a single sink version of Problem 2. It was resurrected recently by Silver [8] in a review paper on production smoothing. Some restrictive assumptions of Bowman's model which Silver finds to be limiting in production smoothing applications are eliminated by our approach.

$q+1$. Hence, if there is an upper bound on the inventory the resulting bound on a partial row and column sum makes it very difficult, if not impossible, to use a transportation algorithm.

By reformulating Problem 2 as a capacitated transshipment or minimum cost flow problem, upper bounds on the inventories can easily be included. In addition, the computer storage requirement for Bowman's approach increases as T^2, and hence the problem size grows very rapidly as the number of periods increases. The computer storage requirement for the transshipment formulation will be seen to increase linearly in T. Furthermore, by recognizing that the transshipment formulation is simply a nested set of transportation problems joined together by inventory variables (arcs), we can decompose it and thereby reduce the storage requirement to approximately the same as for one of the single period transportation problems.

Although for simplicity we limit our presentation only to sink inventories, the algorithm can, in an obvious way, be extended to handle source inventories as well. Also, for simplicity, we shall assume that the shipping variables are uncapacitated; the modification required for capacitated shipments are apparent from the transshipment algorithm. Since convex costs are readily incorporated into a transshipment algorithm, Hu [5], there is an obvious extension to the multiperiod problem with convex costs.

TRANSSHIPMENT FORMULATION

Let A be the set of nodes $A_1^1, A_2^1, \ldots, A_m^T$. Associated with each A_i^t there is a supply a_i^t, i.e., if $x=A_i^t$, let $a(x)=a_i^t$. Similarly, let B be the set of nodes $B_1^1, B_2^1, \ldots, B_n^T$. Associated with each node B_j^t is a demand b_j^t; i.e., if $x=B_j^t$, let $b(x)=b_j^t$.

The set of arcs W contains arcs related to shipments from a source to a sink within the same time period, and arcs representing inventories carried at the sinks. Specifically if $x=A_i^t$ and $y=B_j^t$, then $(x, y)\epsilon W$ for all i, j, and t, and if $x=B_j^{t-1}$, $y=B_j^t$, then $(x, y)\epsilon W$ for all j and $t=2, \ldots, T$.

Let $c(x, y)$ be the unit cost of flow in arc (x, y), $k(x, y)$ the capacity of arc (x, y), and $f(x, y)$ the flow from x to y. Then

$$c(x, y)=\begin{cases} C_{ij}^t \text{ for } x=A_i^t,\ y=B_j^t \\ d_j^t \text{ for } x=B_j^{t-1},\ y=B_j^t \end{cases}$$

and

$$k(x, y)=\begin{cases} \infty \text{ for } x=A_i^t,\ y=B_j^t \\ N_j^t \text{ for } x=B_j^{t-1},\ y=B_j^t. \end{cases}$$

Let

$$f(x, N)=\begin{cases} \sum_{j=1}^{n} X_{ij}^t \text{ for } x=A_i^t \\ I_j^{t+1} \text{ for } x=B_j^t \end{cases}$$

and

$$f(N, x)=\begin{cases} 0 \text{ for } x=A_i^t \\ \sum_{i=1}^{m} X_{ij}^t+I_j^t \text{ for } x=B_j^t. \end{cases}$$

Problem 2 can be rewritten as

minimize $$\sum_{W} c(x, y)f(x, y),$$

subject to

$$f(x, N) \leq a(x) \qquad x\epsilon A$$

$$f(x, N)-f(N, x) \leq -b(x) \qquad x\epsilon B$$

$$0 \leq f(x, y) \leq k(x, y) \qquad (x, y)\epsilon W,$$

which is a capacitated transshipment or minimum cost flow problem with a very particular structure.

As is customary, we add an artificial source node s^* and an artificial sink node t^*. We add arcs (s^*, x) for $x \epsilon A$ with $c(s^*, x)=0$ and, for $x=A_i^t$, define $k(s^*, x)=a_i^t$. Finally, we add arcs (y, t^*) for $y \epsilon B$ and, for $y=B_j^t$, define $c(y, t^*)=h_t$ and $k(y, t^*)=b_j^t$, where $h_1=0$ and

$$h_t=\sum_{p=1}^{t-1}\sum_{i=1}^{m}\sum_{j=1}^{n} C_{ij}^t,$$

for $t=2, 3, \ldots, T$. Usually one defines $c(y, t^*)=0$; however, since $f(y, t^*)=k(y, t^*)$ at termination, the additional cost is a constant

$$H=\sum_{t=1}^{T}\sum_{j=1}^{n} h_t b_j^t.$$

The costs h_t have been chosen to make the cost of satisfying a unit of demand in period t_2 at least as costly as satisfying a unit of demand in period t_1, for $t_1 < t_2$. The reason for doing this is to force the algorithm to make as many flow changes in period t as possible before it considers flow changes in period $t+1$. In fact, the algorithm will satisfy all demand in period t_1 before any demand in period $t_2 > t_1$ is satisfied. As will be seen, this is highly desirable in the decomposition approach, since it tends to minimize the number of data transfers from auxiliary to core storage.

The transshipment formulation of problem 2 is shown in Figure 1 for $m=n=T=2$, where the parentheses indicate (cost, capacity) of the specified arc.

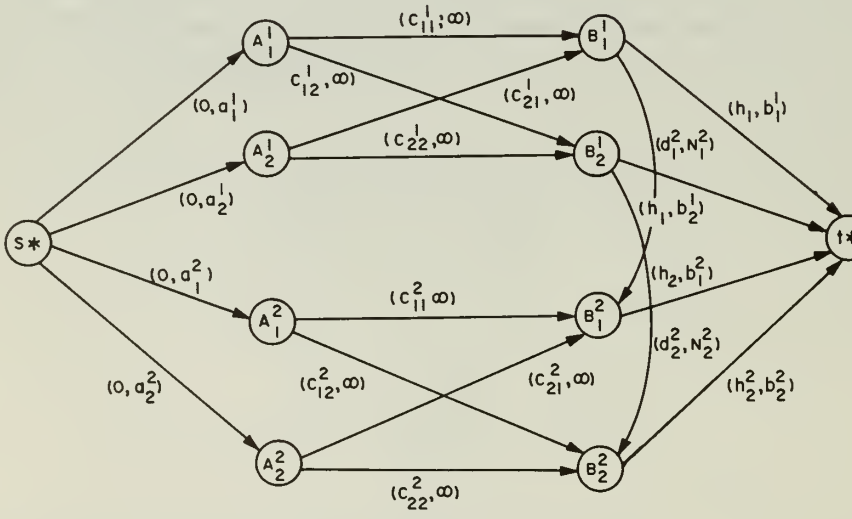


Figure 1

ALGORITHM

The algorithm is a standard algorithm for the transshipment problem with the labelling procedures suitably modified to take advantage of the structure of the network. We shall describe it in terms of the Busacker and Gowen [2] algorithm (also see Hu [6]), although it could be implemented within the framework of the Ford and Fulkerson [4] algorithm.

The Busacker-Gowen algorithm consists of solving a sequence of shortest path problems from the source s^* to the sink t^* (see Figure 1). We first present the algorithm in its most straight forward form, a form whose computer storage requirement increases linearly with T. The distances, along the arcs, which are used to determine the shortest path are functions of existing flow in the network. We begin

with $f(x, y) = 0$ for all arcs. Each directed arc, (x, y) is thought of as an undirected edge with distance given by

$$c^*(x, y) = \begin{cases} c(x, y) & \text{if } 0 \leq f(x, y) < k(x, y) \\ \infty & \text{if } f(x, y) = k(x, y) \end{cases}$$

and

$$c^*(y, x) = \begin{cases} -c(x, y) & \text{if } f(x, y) > 0 \\ \infty & \text{if } f(x, y) = 0. \end{cases}$$

We find a shortest path from s^* to t^* and then ship the maximum amount possible (i.e., maintaining $0 \leq f(x, y) \leq k(x, y)$) along that path. When flows are increased the distances are revised and the shortest path recalculated. The algorithm terminates when all demands have been satisfied.

Although some distances are negative, we never have negative cycles since the flow pattern is a minimal cost pattern at every iteration. Thus, any shortest path algorithm which allows negative distances can be used.† We will describe one shortest-route algorithm, so that we may discuss a modification of the Busacker-Gowen method that is useful for our problem.

A node is in one of two states, scanned or unscanned. All nodes are labelled, a label is an ordered pair $(L(x), p(x))$; at termination, $L(x)$ is the length of a shortest path from s^* to x, and $p(x)$ is used to "decode" the shortest path. Initially s^* is unscanned and is labelled $(L(s^*) = 0,\ p(s^*) = -1)$. All other nodes are scanned and labeled $(L(x) = \infty,\ p(x) = 0)$.

We chose an unscanned node, say x, and compare $L(x) + c^*(x, y)$ with $L(y)$ for all nodes y which are adjacent to node x. If $L(x) + c^*(x, y) < L(y)$, we replace $L(y)$ with $L(x) + c^*(x, y)$, set $p(y) = x$, and consider node y unscanned. When all nodes adjacent to x have been tested, node x is considered scanned. This procedure is repeated until there are no unscanned nodes.

If we were dealing with an arbitrary transshipment problem, it would be necessary to have all nodes scanned before initiating a flow change. After each flow change, we would have to discard all labels and restart the algorithm with $L(s^*) = 0$ and $L(x) = \infty$ for $x \neq s^*$. Because of the structure of our problem, we can do better.

Notice from Figure 1 that the only connections between one period and the next are the arcs representing inventory flows (I_j^t), and excluding these arcs, the network represents a series of T independent transportation problems. This structure will be used to modify the implementation of the transshipment algorithm.

When scanning a sink node, say B_j^t, it will be convenient to speak of three different kinds of scanning operations; in period scanning, backward scanning, and forward scanning. By in period scanning, we mean either attempting to label a source A_i^t from B_j^t, or attempting to label t^*; backward scanning refers to attempting to label B_j^{t-1}, from B_j^t; and forward scanning refers to attempting to label B_j^{t+1} from B_j^t.

Assuming that $C_{ij}^t \geq 0$, then from our definition of the arc cost h_t, it is apparent that a shortest path (i.e., a minimal cost flow augmenting path) will not involve nodes of period t until all demands of period t' have been satisfied, for all $t' < t$. Suppose the demands for all sinks of period t have been satisfied for $t < t_2$, and suppose there exists at least one sink $B_j^{t_2}$ with unsatisfied demand. Then when scanning node s^*, we only test nodes A_i^t for $t \leq t_2$. Further, when scanning a sink node $B_j^{t_2}$, we can eliminate forward scans.

†Two recent review papers describing such shortest path algorithms are Dreyfus [3] and Hu [6].

Now consider the labels that must be discarded after a flow change that involved periods t_1 through t_2. If this flow change has just satisfied all of the demands in period t_2, then redefine t_2 as t_2+1. Let $t_0 < t_1$ be the largest t with no labels from period t_0+1. After a flow change, one need only discard the labels in periods t_0+1 through t_2 (i.e., set $L(x)=\infty$ and consider the node scanned). In addition, the label on t^* is discarded and if $t_0 > 0$, then nodes $B_k^{t_0}$, $k=1, 2, \ldots, n$ are considered unscanned, since forward scanning from these nodes is required. Finally, of course, one must consider s^* as unscanned.

In the simplest case when the quantity of flow added was uniquely determined by the unsatisfied demand at $B_j^{t_2}$ only the label on t^* needs to be discarded and nodes $B_k^{t_2}$, for $k=1, 2, \ldots, n$ are considered unscanned.

If the data for all periods do not fit into rapid access storage, it is possible to implement the algorithm with only the data for one period at a time in core. We reserve sufficient rapid access storage to handle the data for $(m+2n+2)$ nodes and $(mn+m+2n)$ arcs. In dealing with period t, the data in rapid access storage is for nodes s^*, t^*, A_i^t, B_j^t and B_j^{t-1}, for all i and j, and for acrs (s^*, A_i^t), (A_i^t, B_j^t), (B_j^t, t^*) and (B_j^{t-1}, B_j^t), for all i and j. Note that a standard transportation problem would require storage for $(m+n+2)$ nodes, s^*, t^*, A_i, and B_j, for all i and j; and $(mn+m+n)$ arcs, (s^*, A_i), (A_i, B_j), and (B_j, t^*), for all i and j.

Initially all flows are zero, and we deal exclusively with first period data (i.e., initially all flow changes are to satisfy demands in period 1). Consider the general step of the algorithm; that is, suppose we have just completed a flow change and t_0 and t_2 have been defined. Set $L(t^*)=\infty$ and $L(s^*)=0$; bring the data for period $t=t_0+1$ into rapid access memory and set $L(x)=\infty$ for $x=A_i^t$ and B_j^t for all i and j. If $t_0 > 1$, bring in the current value of $L(B_j^{t-1})$, for all j. Nodes t^*, A_i^t and B_j^t, for all i and j are considered scanned; nodes s^*, and B_j^{t-1} (if $t_0 \geq 1$) are unscanned. Now proceed with a shortest route problem, except do not attempt forward scanning to nodes of type B_j^{t+1}. If $L(B_j^{t-1})$ is redefined for any j we set a "flag" (this cannot happen for period $t=t_0+1$). When we complete all scanning possible in period t, the flag is checked, if it is set, we replace the data in rapid access memory with the data for period $t-1$ and repeat the above process. If the flag is not set, and if $t < t_2$, we replace the data in rapid access memory with the data for period $t+1$ and repeat the above process. Eventually, we will complete all labelings of period t_2. At this point, we use the shortest path found to t^* and increase the flow.

EXAMPLE

The data for a two-source, two-sink, two-period problem is given in Figure 2. The ordered triples (c, k, f) on each arc refer to cost, capacity and flow, respectively. The reverse arcs, created by positive flows are not explicitly shown. Note that $h_2=c_{23}+c_{14}=2$. While one usually begins with $f(x, y)=0$ for all arcs, the two obvious zero cost flows of three units each with the use of the chains $(s, 1, 3, t)$ and $(s, 2, 4, t)$ are included on Figure 2.

The labeling procedure is now begun. The ordered pairs $(L(x), p(x))$ adjacent to each node are the labels at termination of the first shortest-path calculation. The shortest path of length 12 is given by the chain $(s, 2, 3, 7, t)$. The flow increase on this path is one unit. The adjusted costs, flows, and new labels are given in Figure 3 (all labels must be recalculated). The shortest path of length 13 is given by the chain $(s, 5, 7, t)$ and the flow increase on this path is five units. Although this flow change only involves period 2, all labels must be discarded since there is reverse labeling from period 2 to period 1. Two more flow changes are required—one unit on the path $(s, 6, 7, 3, 2, 4, 8, t)$, with decreases

of one unit on arcs $(7, 3)$ and $(3, 2)$, and four units on the path $(s, 6, 8, t)$. The optimal flow is shown in Figure 4.

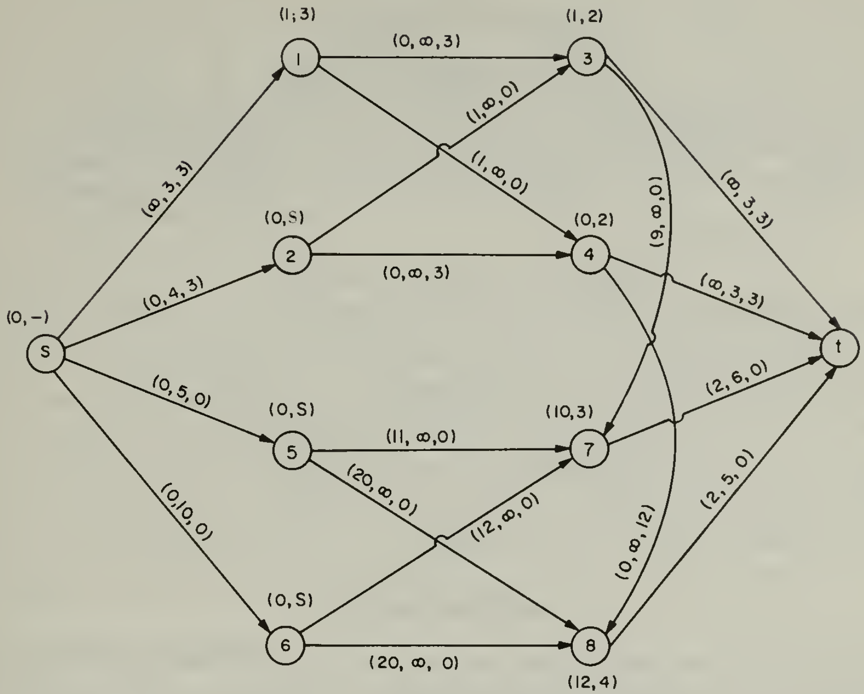


Figure 2

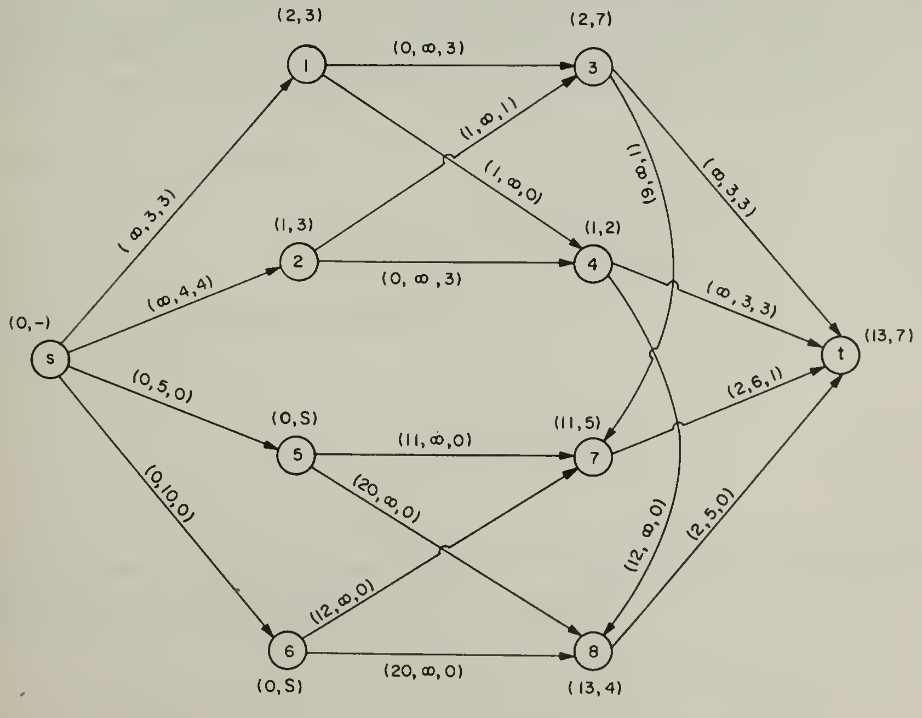


Figure 3

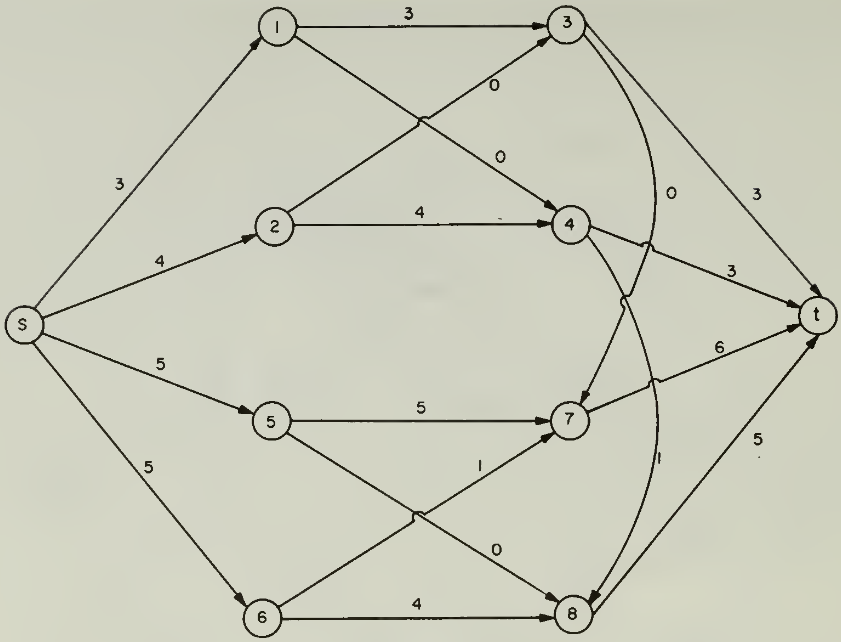


Figure 4

REFERENCES

[1] Bowman, E. R., "Production Scheduling by the Transportation Method of Linear Programming," Operations Research, *4*, 100–103 (1956).

[2] Busacker, R. G. and P. J. Gowen, "A Procedure for Determining a Family of Minimal Cost Network Flow Patterns," Operations Research Office, Tech. Paper 15 (1961).

[3] Dreyfus, S. E., "An Appraisal of Some Shortest Path Algorithms," Rand Memo RM–5433–PR (1967).

[4] Ford, L. R. and D. R. Fulkerson, *Flows in Networks* (Princeton University Press, Princeton, N.J., 1963).

[5] Hu, T. C., "Minimum Convex-Cost Flows in Network," Rand Memo RM–4265–PR (1964).

[6] Hu, T. C., "Recent Advances in Network Flows," Mathematics Research Center Report #753, University of Wisconsin (1967).

[7] Kreibel, C. H., "Warehousing and Transshipment Under Seasonal Demand," J. Regional Science, *3*, 57–71 (1961).

[8] Silver, E. A., "A Tutorial on Production Smoothing and Work Force Balancing," Operations Research, *15*, 958–1011 (1967).

* * *

ON THE EQUIVALENCE BETWEEN THE NODE-ARC AND ARC-CHAIN FORMULATIONS FOR THE MULTI-COMMODITY MAXIMAL FLOW PROBLEM

John J. Jarvis

Georgia Institute of Technology

ABSTRACT

The purpose of this article is to formulate the multi-commodity maximal flow problem into a node-arc form and to show that when decomposition is applied to this form the resulting master and subproblems become precisely those described by Ford & Fulkerson [3] using the arc-chain formulation. A generalization to the problem is then considered which can potentially speed its convergence.

INTRODUCTION

Recently, an article by J. A. Tomlin [6] formulated the minimum cost multi-commodity flow problem in node-arc form as a linear programming problem. In his formulation, the flow of each commodity was specified and, therefore, was not a variable in the problem. By a similar procedure one could formulate the multi-commodity maximal flow problem in node-arc form as a linear programming problem (in which the amount of each commodity is unspecified). Tomlin has shown that the node-arc and arc-chain formulations for the minimal cost flow problem are equivalent, within a change of variable. We shall show the equivalence between the two formulations for the maximal flow problem does not require a change of variable, a fact derivable from the nature of the subproblems.

NODE-ARC FORMULATION

The multi-commodity maximal flow problem can be formulated as follows. Let $G[N;A]$ be a network with n nodes and m directed arcs. Associated with each arc (i, j) is a capacity $c_{ij} \geqq 0$. Further, assume there are r commodities with a single source, s_k, and a single sink, t_k, for the kth commodity.

We, using a notation similar to Tomlin's, denote by y_{ij}^k the flow of commodity k on arc (i, j). The arc capacity constraints become

(1) $$\sum_{k=1}^{r} y_{ij}^k \leqq c_{ij} \qquad (i, j = 1, \ldots, n).$$

The node conservation equations for each commodity are

(2) $$\sum_{j=1}^{n} (y_{ij}^k - y_{ji}^k) = \begin{cases} v_k \text{ for } i = s_k \\ -v_k \text{ for } i = t_k \\ 0 \text{ otherwise.} \end{cases}$$

In matrix form, the conservation equations are

(2a) $$\mathbf{A}\mathbf{Y}_k = -\mathbf{d}_k v_k$$

or

$$[\mathbf{d}_k, \mathbf{A}] \begin{bmatrix} v_k \\ \mathbf{Y}_k \end{bmatrix} = \underline{\mathbf{0}},$$

where $\mathbf{A}$ is the node-arc incidence matrix for the network, $-\mathbf{d}_k$ is the vector of coefficients for v_k in Eq. (2), and $\mathbf{Y}_k$ is the vector of flows for the kth commodity.

The node-arc formulation of the multi-commodity maximal flow problem then becomes

$$\text{Max } [1, \mathbf{0}]\begin{bmatrix} v_1 \\ \mathbf{Y}_1 \end{bmatrix} + \ldots + [1, \mathbf{0}]\begin{bmatrix} v_r \\ \mathbf{Y}_r \end{bmatrix} \tag{3}$$

$$\text{s.t. } [0, \mathbf{I}]\begin{bmatrix} v_1 \\ \mathbf{Y}_1 \end{bmatrix} + \ldots + [\mathbf{0}, \mathbf{I}]\begin{bmatrix} v_r \\ \mathbf{Y}_r \end{bmatrix} + \mathbf{IS} = \mathbf{C}$$

$$[\mathbf{d}_1, \mathbf{A}]\begin{bmatrix} v_1 \\ \mathbf{Y}_1 \end{bmatrix} = \mathbf{0}$$

$$\vdots$$

$$[\mathbf{d}_r, \mathbf{A}]\begin{bmatrix} v_r \\ \mathbf{Y}_r \end{bmatrix} = \mathbf{0}$$

$$\mathbf{S} \geqq 0,\ \mathbf{Y}_k \geqq 0 \text{ for } k = 1, \ldots, r,$$

where $\mathbf{S}$ is a vector of slack variables for each arc and $\mathbf{I}$ is the unit matrix.

DECOMPOSITION

We are able to utilize the Dantzig-Wolfe decomposition principle [2], because of its form, to reduce the size of the above linear program. Before applying decomposition, we call attention to the special property possessed by the constraints (2a) which will be the constraints for each subproblem. Equations (2a) are a set of homogeneous equations and all solutions of (2a) are called homogeneous solutions. We say that $\mathbf{X}$ is a homogeneous solution to a set of equations if $\rho\mathbf{X}$ for all $\rho \geqq 0$, is also a solution to the set of equations. It is easily seen that all solutions of (2a) have the homogeneous property. The direct consequence of this homogeneous property is that the convex set defined by (2a) contains only extreme rays. Thus, even though we are not able to define each point in the convex set of (2a) by a convex combination of extreme points we are able to define the point by a non-negative combination of the homogeneous solutions defining the extreme rays of the set.

Let $[v_{kj}^*, \mathbf{Y}_{kj}^*]^T$, $j = 1, \ldots n_k$ be all of the homogeneous solutions to (2a). Then we may write

$$\begin{bmatrix} v_k \\ \mathbf{Y}_k \end{bmatrix} = \sum_{j=1}^{n_k} \rho_{kj} \begin{bmatrix} v_{kj}^* \\ \mathbf{Y}_{kj}^* \end{bmatrix}, \tag{4}$$

where

$$\rho_{kj} \geqq 0 \text{ for all } j = 1, \ldots, n_k.$$

The original problem, through decomposition, becomes

$$\text{max. } \sum_{j=1}^{n_1} \rho_{1j} [1, \mathbf{0}] \begin{bmatrix} v_{1j}^* \\ \mathbf{Y}_{1j}^* \end{bmatrix} + \ldots + \sum_{j=1}^{n_r} \rho_{rj} [1, \mathbf{0}] \begin{bmatrix} v_{rj}^* \\ \mathbf{Y}_{rj}^* \end{bmatrix} \tag{5}$$

$$s.t. \sum_{j=1}^{n_1} \rho_{1j} [\mathbf{0}, I] \begin{bmatrix} v_{1j}^* \\ Y_{1j}^* \end{bmatrix} + \ldots + \sum_{j=1}^{n_r} \rho_{rj} [\mathbf{0}, \mathbf{I}] \begin{bmatrix} v_{rj}^* \\ Y_{rj}^* \end{bmatrix} + IS = C$$

$$\rho_{kj} \geqq 0. \text{ all } k = 1, \ldots, r, j = 1, \ldots, n_k.$$

This is a problem with m equations and a large number of variables.

THE SUBPROBLEMS

We begin with the basic feasible solution in which each of the slacks is equal to the arc capacity and all commodity and arc flows are equal to zero. After assigning simplex multipliers $\pi_1, \ldots, \pi_m$ to the m equations of (5), we insert ρ_{kj} into the basis if

$$[1, \mathbf{0}]\begin{bmatrix} v_{kj}^* \\ \mathbf{Y}_{kj}^* \end{bmatrix} - \boldsymbol{\pi}_B[\mathbf{0}, \mathbf{I}]\begin{bmatrix} v_{kj}^* \\ \mathbf{Y}_{kj}^* \end{bmatrix} > 0$$

or

$$v_{kj}^* - \boldsymbol{\pi}_B Y_{kj}^* > 0,$$

where $\boldsymbol{\pi}_B = (\pi_1, \pi_2, \ldots, \pi_m)$. The method used to find ρ_{kj} is to solve the subproblem

$$\max v_k - \boldsymbol{\pi}_B \mathbf{Y}_k$$

$$s.t.\ [\mathbf{d}_k, \mathbf{A}]\begin{bmatrix} v_k \\ \mathbf{Y}_k \end{bmatrix} = \mathbf{0}$$

$$\mathbf{Y}_k \geqq \mathbf{0}$$

This subproblem is a minimum cost flow problem on an uncapacitated network. The solution to this subproblem is to find the shortest chain from s_k to t_k using π_i, $i = 1, \ldots, m$ as the arc weights. We note that all $\pi_i \geqq 0$, for, if not, we can insert the appropriate slack variable in the basis and obtain an increase. Then if the sum of the π_i along the shortest chain is less than one, the solution is to place an infinite amount of flow along this chain.

The solution to the subproblem is unbounded, as could be expected from the homogeneous property associated with (2a). Since it is only necessary to have the form of the extreme ray, we may take the homogeneous solution to the subproblem produced when one unit of flow is placed along the shortest chain. Then, the homogeneous solution vector becomes

$$\begin{bmatrix} \mathbf{v}_{kj}^* \\ \mathbf{Y}_{kj}^* \end{bmatrix} = \begin{bmatrix} 1 \\ \mathbf{a}_{kj} \end{bmatrix},$$

where $a_{kj} = [a_{ij}]$ is a vector of zeros and ones defined by

$$a_{ij} = \begin{cases} 1 \text{ if arc } (i, j) \text{ is in the shortest chain} \\ 0 \text{ otherwise.} \end{cases}$$

Finally, the ρ_{kj} activity vector is

$$[\mathbf{0}, \mathbf{I}]\begin{bmatrix} 1 \\ \mathbf{a}_{kj} \end{bmatrix} = \mathbf{a}_{kj},$$

after which one inserts ρ_{kj} into the basis to obtain a new basic feasible solution, thereby completing one iteration of the decomposition procedure.

One immediately recognizes that this is precisely the algorithm derived by Ford and Fulkerson [3] in considering the arc-chain formulation of the multi-commodity maximal flow problem. Further, we see that ρ_{kj} is the amount of flow that is allocated to the chain defined by $\mathbf{a}_{kj}$.

The conclusion, then, is that it makes no difference whether one uses the node-arc or arc-chain formulation for the multi-commodity maximal flow problem since the two methods are precisely the same. In fact, Dantzig [2] indicates that Ford & Fulkerson [3] motivated decomposition.

GENERALIZATION

In the foregoing, we saw that at each iteration of the two algorithms one additional chain was being added to the basis. There are ways of admitting more than one chain to the basis at each iteration if one generalizes the node-arc formulation. This has the potential effect of speeding up the convergence of the algorithm at a cost of additional effort to solve the subproblems.

The generalization is as follows. We notice that the constraints defined by (1) could also be accompanied by

$$(6) \qquad y_{ij}^k \leqq c_{ij} \qquad (k=1,\ldots,r;\, i,j=1,\ldots,n).$$

In this case, the subproblem becomes

$$\begin{aligned} &\max.\ v_k - \boldsymbol{\pi}_B \mathbf{Y}_k \\ &\text{s.t.}\ \ [\mathbf{d}_k, \mathbf{A}]\begin{bmatrix} v_k \\ \mathbf{Y}_k \end{bmatrix} = \underline{\mathbf{0}} \\ &\qquad \mathbf{0} \leqq \mathbf{Y}_k \leqq \mathbf{C} \end{aligned}$$

and is a minimum cost flow problem on a capacitated network. One can use the algorithms of Ford and Fulkerson [4], Busacker and Gowen [1], or Klein [5] to obtain a solution, in general, comprised of many chains. The master problem remains the same as (4) and (5). This results from the fact that

$$\sum_{k=1}^{r} y_{ij}^k \leq c_{ij} \Rightarrow y_{ij}^k \leq C_{ij},$$

i.e. any feasible solution to the master is also feasible for the subproblems.

As an example of the potential power of the generalized method consider the network in Figure 1. The shortest chain procedure required four iterations (inversions of a 7×7 basis matrix using the revised simplex method) to achieve the optimal solution.

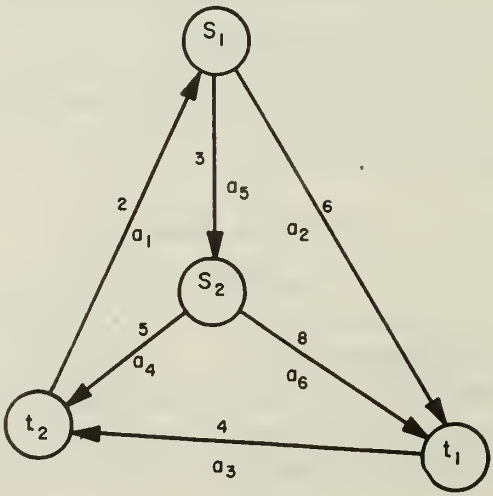


Figure 1

Applying the generalized procedure to the graph of Figure 1, we begin with $\pi_B = (0, 0, 0, 0, 0, 0)$ and obtain $\mathbf{a}_{11} = (-9, 0, 6, 0, 0, 3, 3)^T$. Inserting $\mathbf{a}_{11}$ into the basis yields $\rho_{11} = 1$ and $\pi_B = (0, 0, 0, 0, 3, 0)$. This new set of dual variables gives $\mathbf{a}_{21} = (-9, 0, 0, 4, 5, 0, 4)^T$. Inserting $\mathbf{a}_{21}$ into the basis we obtain $\rho_{11} = 1$, $\rho_{21} = 1$, and $\pi_B = (0, 0, 9/4, 0, 3, 0)$, which is optimal. Thus, we have two iterations instead of the previous four.

More effort is required to solve a minimum cost flow problem than a shortest chain problem; however, a minimum cost flow problem produces more chains in the ρ_{kj} activity vector which requires potentially less pivots to achieve optimality. The trade-off becomes roughly, the effort to transform the basis matrix and solve the associated shortest chain problems several times against the effort to solve a minimum cost flow problem.

It should also be noted that this same set of generalizations applies, in a slightly different manner, to the minimum cost flow problem described by Tomlin. In this case the node representing sink, t_k, is replaced with two nodes, t_k' and t_k'', in the manner described by Ford & Fulkerson [4] with the capacity of the new arc (t_k', t_k'') equal to the amount of flow required and large negative cost. We then proceed to apply the algorithm with new objective and with the subproblems being minimal cost flow problems between s_k and t_k'' on a capacitated network. Again, it is only necessary to have one constraint in the master for each arc.

REFERENCES

[1] Busacker, R. G. and P. J. Gowen, "A Procedure for Determining a Family of Minimal-Cost Flow Patterns," Operations Research Office Technical Paper 15 (1961).

[2] Dantzig, G. B., *Linear Programming and Extensions* (Princeton University Press, Princeton, N.J., 1963).

[3] Ford, L. R. and D. R. Fulkerson, "A Suggested Computation for Maximal Multi-Commodity Network Flows," Management Science **5**, 97–101 (1958).

[4] Ford, L. R., and D. R. Fulkerson, *Flows in Networks* (Princeton University Press, Princeton, N.J., 1962).

[5] Klein, M., "A Primal Method for Minimal Cost Flows With Applications to the Assignment and Transportation Problems," Management Science **14**, 205–220 (1967).

[6] Tomlin, J. A., "Minimum-Cost Multicommodity Network Flows," Operations Research **14**, 45–51 (1966).

* * *

QUANTIFICATION OF CONTRACTOR RISK

Clifford W. Marshall
Polytechnic Institute of Brooklyn

ABSTRACT

This paper presents a quantitative index of the level of risk assumed by a contractor in various contract type situations. The definition includes expression of real world uncertainty and contractor's utility for money. Examples are given for the major contract types and special applications are discussed.

INTRODUCTION

A contractor is an entrepreneur who operates within constraints imposed by a contract when providing goods or services to its customer. A contractor's risk is the possibility of undesirable variation in its fee due to factors having uncertain future outcomes. Risks assumed by a contractor differ from general entrepreneural risk in that contract structure tends to limit the risk a contractor may assume. Moreover most contracts provide direct payment to the contractor for its assumption of risk. The general entrepreneur may choose to operate in a highly risky environment and to earn whatever subsequent events allow. If a contractor's risk is to be limited and it is to be compensated for whatever risk it does assume, a quantitative measure of such risk is highly desirable. Lacking such a measure the risk element must be treated in a purely subjective fashion as indeed is the common practice. A major difficulty with that practice is the lack of clear expressions of risk assumption that can be compared for different situations, directly represented for discussion purposes, or numerically defended against alternative options. In defining a quantitative measure for contractor risk assumption the subjective element occurs in the definition itself and in its interpretation; however, the subjective element is contained in the quantification theory rather than in particular details of individual situations.

This paper presents a quantification of contractor risk. Other quantifications are certainly possible and one should not say that a particular quantification is true or not true, but rather that it is useful or not useful. The quantification defined here will be shown to be based on intuitively desirable features and its utility will be demonstrated by various applications. Contractor risk will be quantified for each major contract type. Risks will be compared for different situations and application will be made to contract type selection and to structuring. Additional features of contractor risk assumption will be clarified.

DEFINITION OF RISK QUANTIFICATION

Risk is introduced into contractual situations by the existence of real world uncertainties. Such uncertainty may be produced by numerous causes resulting in the necessity for random descriptions of future results. The risk in such situations is produced by the spread or possible variations of future events. If there was only one known event there would be no risk even if the event was undesirable. The less extreme situation in which one must choose between a desirable expected event subject to wide variation and a less desirable expected event with little variation (and therefore little risk) is very

common in real world situations. Many aspects of a contractual situation may be subject to the element of risk due to the presence of unknown future events. This paper will deal with the risk associated with the cost of a contracted item. Other aspects such as schedule or performance may be studied by the same methods. Combination of the various risks that one might quantify in a contract represents a generalization of the ideas presented here that will not be considered as part of this paper.

To formulate a measure of contractor risk due to variation in cost one wishes to consider three major factors. The variation in cost due to real world uncertainty should certainly be a major factor. Contract structure (including contract type) also plays a major role in contractor risk assumption. In addition, contractors with extensive resources or special goals may treat money in a very different way than other contractors do. Thus contractor utility for money should be included in the description of its level of risk assumption. All these factors will be included in our quantification of contractor risk assumption.

The variance of a random quantity is a convenient measure of variability. It has the requisite intuitive properties,* and has a relatively simple mathematical formulation.† Accordingly the variance of an appropriate random quantity will be introduced here as a measure of contractor risk assumption. The random quantity employed will be a composite of the three factors—cost variability, contract structure, and contractor's utility for money—that are basic to a representation of contractor risk.

In order to formulate the quantitative measure of contractor risk in mathematical terms, the following notation is required.

Let C denote a random variable whose values represent final cost to the customer of a contracted item. In specific analyses of contractor risk, it would be assumed that the probability distribution govering C was known. Various forms may be considered for such distributions as will be discussed subsequently.**

Let $F(c)$ be the contractor's fee as specified by the contract when the contracted item has final cost $C=c$. The function $F(\cdot)$ is a mathematical expression of contract structure. It assumes various forms characteristic of the different contract types. Several forms will be used subsequently to quantify contractor risk for each of the major contract types.

Let $u(\cdot)$ represent a contractor's utility function for money. The most common cases are where $u(\cdot)$ is linear, quadratic, or exponential in form. In specific analyses of contractor risk, $u(\cdot)$ will be assumed to have a specific form.††

The quantification of contractor risk is now expressible in terms of the quantities defined above. If the measure of risk is denoted by R it is defined by the following formula:

(1) $$R=\operatorname{Var}\left[u(F(C))\right].$$

Thus if $G_C(c)$ is the distribution function for the random cost C, R may be expressed as the Stieltje's integral form:

(1) $$R=\int_c \left[u(F(c))-\bar{u}\right]^2 dG_C(c),$$

where $$\bar{u}=\int_c u(F(c))\,dG_C(c).$$

*Variance includes both positive and negative variation about a central value and weights larger variations more heavily than small variation.

†Variance is simple as compared with expected values of absolute magnitude or higher even moments of variation about a central value.

**For the purposes of this paper the only theoretical requirements on the probability distribution is that C have finite expected value and finite variance.

††In practice it may be difficult to specify $u(\cdot)$ in exact form for a specific contractor; however, one can often gain understanding about problems by performing analyses in which reasonable mathematical forms for $u(\cdot)$ are assumed.

The quantification of contractor risk defined by Eq. (1) derives from the basic observation that risk depends on variation in item cost. That variation, however, must be modified by contractual constraints that limit the amount of assumption of risk by the contractor. Definition (1) also includes the effect of contractor utility for money so that the level of risk assumption may be based on a contractor's values rather than directly on money. The quantity R is seen to satisfy intuitively desirable properties for a measure of contractor risk. R represents variation in the utility value of contractor fee and therefore measures contractor risk. Though variation in fee allows increased fee as well as less fee, sound management will wish to minimize such variation when due to factors that are largely outside management control. A risk capital entrepreneur may desire high risk situations in which the large variation of possible outcomes allows the possibility of large gain (when uncertain favorable future events occur). Normal corporate management limited variation situations in which decisions can be made with high confidence as to expected future results. Failure to have limited variation is considered as risk by management.

In the next section values assumed by R in different contract situations will be given. These will show that R has a desirable form relative to the levels of contractor risk assumption commonly attributed to each major contract type. It remains to present in this section a discussion of possible interpretations for R.

There are at least two distinct points of view regarding interpretation of R. One may consider R strictly as a measure of risk assumption and compare that measure for different situations to indicate which situations correspond to greater assumption of contractor risk. In such an interpretation numerical values of R have little or no specific meaning relative to contract parameters (such as target fee). This is the most direct and common use of a measure. It allows relative comparison of magnitudes without possessing any absolute meaning. The possibility exists for another interpretation of R in terms of contractor fee. In order to explore this interpretation it will be necessary to discuss the factors comprising the concept of contractor fee.

Contractor fee may be considered to be the sum of three distinct types of fees which may be called service fee, actuarial fee, and incentive fee. Service fee is the basic amount paid to the contractor for doing work, it may be considered to act as a corporate salary analogous to the salary paid to an individual for doing work. Actuarial fee is the compensation paid to a contractor for its assumption of risk. Incentive fee is paid for deviations from a target (standard) product under contract. A positive incentive fee is a reward for improved results and a negative incentive fee is imposed as a penalty for less than target (standard) achievement. The concepts of incentive fee and actuarial fee are often closely related, but should be distinguished. A sharp decrease in incentive fee might be due to the element of uncertainty rather than to factors directly subject to contractor responsibility. The presence of such uncertainty in a contract situation providing for sharp decrease in incentive fee should be compensated to some degree by an actuarial fee. For a fixed slope of the share line in an incentive contract the target cost will determine actual fee values. When above target costs are possible due to uncertainty (i.e., in a risk situation) the fee values should be increased by an appropriate amount of actuarial fee (added to target fee). In the absence of such uncertainty the actual fee should be less. This situation is illustrated in Figure 1 for a cost plus incentive fee contract.* Figure 1(a) corresponds to a higher risk situation than does Figure 1(b). Both contracts have the same structure except for an increased actuarial fee added in case (a) to compensate for the factor of greater risk. Thus it

*Details of contract types are presented in the next section.

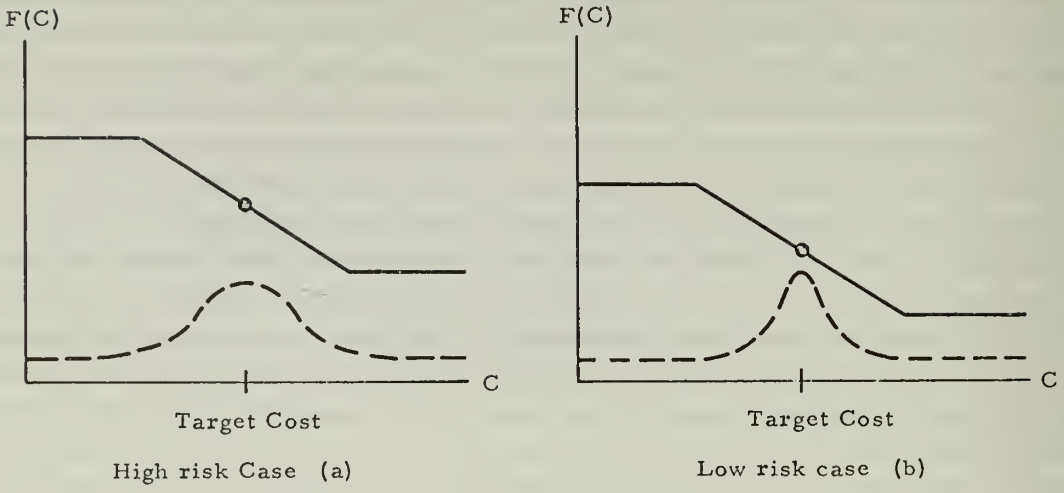


Figure 1. Effect of actuarial fee.

may be seen that there is an intricate trade off between actuarial fee and the sensitivity of incentive fee to uncertainty (usually indicated by the slope of the share line in incentive contracts) that is often accommodated subjectively in the target fee. That accommodation is probably accomplished without any conscious realization of the two types of fees as distinct and certainly without quantitative measurement of actuarial fee. The possibility seems to exist for using R and the contract structure on which it is defined to clarify the relative importance of incentive and actuarial effects and to help arrive at quantitative assessment for reasonable actuarial fee. Such a result would allow the composition of target fee from service and actuarial fees to be less subjectively arrived at. It would also clarify the ambiguous, but widely made, remarks to the effect that changes in incentive fee indicate contractor's assumption of risk. It is shown above that this is only true in an indirect way. The incentive should not be paid for risk assumption. Risk assumption should be compensated by a fixed actuarial fee incorporated in the target fee. Such actuarial fee should, however, reflect the slope of the share line in incentive contracts because in those cases variation in incentive fee is certainly part of the result of the contractor's assumption of risk.

RISK ASSUMPTION FOR CONTRACT TYPES

In this section R is expressed for each of the major contract types.* The expressions for R show how well it reflects the levels of risk assumption commonly attributed to each contract type. Some numerical results illustrate the way R behaves as a measure of risk for various levels of uncertainty on cost C.

The definition of R allows any contractor utility function for money, $u(\cdot)$ to be used. For purposes of this paper, it is sufficient to employ a simple linear utility for money of the form $u(x)=x$ for all x. Thus in the formulations to follow, the value of money is taken as equal to money. This simplifies the expressions for R, allowing useful observations to be made relative to the basic formulation of R.

Table 1 gives the forms for R for each major contract type in terms of a general probability density

*Award fee type contracts are not considered in this paper. Such contracts have a completely different structure than the types considered here and include uncertain factors of a very different sort.

$g_C(c)$ for cost C.† Though the expressions could be given in more general form by using probability distributions $G_C(c)$ there is little need to do so. Most situations will have C governed by a probability density function and those situations that do not can be treated in terms of Eq. (1), (which yields sums for discrete random variables C, etc.) to obtain forms for R in cases of interest.

For the benefit of those readers who may not be familiar with the major contract types each will now be briefly described:

TABLE 1. *Risk Assumption for Contract Types* [a]

Contract type	Fee Formula [b]	Measure of Risk [c]	Level of Contractor Risk usually attributed to contract type
FFP	$F(c)=P-c$	$R=\text{Var}(C)$	Full risk assumption.
FPIF (*FPI*)	$F(c)=mc+b, \quad c\leq c_0$ $=F(c_0)-c+c_0, \quad c>c_0$ $mc_0+b+c_0=P_0$ $c_0=PTA$	$R=m^2\int_0^{c_0}(c-c_E)^2g_C(c)dc$ $+\int_{c_0}^{\infty}(c-c_E)^2g_C(c)dc$ $c_E=E(C)$	High risk assumption.
CPIF	$F(c)=mc_0+b, \quad 0\leq c\leq c_0$ $=mc+b, \quad c_0\leq c\leq c_1$ $=mc_1+b, \quad c\geq c_1$	$R=m^2\int_{c_0}^{c_1}(c-c_E)^2g_C(c)dc$	Low risk assumption.
CPFF	$F(c)=F_0$	$R=0$	No assumption of risk.

[a] Utility for money is equal to money in all cases.
[b] The notation is as defined above under the discussion of contract types.
[c] $g_C(c)$ denotes the probability density of C. It may be any density for which first and second moments of C exist. c_E is the expected value of C.

Firm Fixed Price Contract (FFP)

In this type of contract the contractor agrees to provide a product (or service) at a fixed price to the customer. Contractor fee is the difference between the price and the cost (which in general is a random variable C).

Fixed Price Incentive Firm Contract (FPIF)

In contracts of this type there is a price ceiling P_0 such that the customer pays no more than P_0. An incentive fee is provided to motivate the contractor to keep cost down. The most common form of incentive fee is expressed in terms of two straight lines as follows:

$$F(c)=mc+b \text{ for } c\leq c_0$$
$$F(c)=F(c_0)-(c-c_0) \text{ for } c>c_0.$$

The cost value c_0 is known as the point of total assumption (PTA). At that value of cost the contractor must absorb increase in cost out of its fee. For large cost over runs the fee can go negative. In the formulas m is the slope of the share line, b is the (hypothetical) fee value for zero cost, and $P_0=(m+1)c_0+b$. Since one may never achieve zero cost the value of b is often derived from target cost value c_T in which $F(c_T)-mc_T=b$. It may be observed that the contract is specified by the values of target cost, target fee, slope of share line (directly related to the sharing ratio), and ceiling price P_0.

†The forms presented are general and apply to any customer and contractor except for legal restrictions imposed in some cases when the customer is the Government (e.g., DoD or NASA). The forms given contain those restrictions since the Government is by far the most extensive user of these contract types.

Cost Plus Incentive Fee Contract (CPIF)

These contracts are similar to FPIF contracts except there is no price ceiling and a minimum fee is specified. In CPIF contracts there is a maximum fee allowable by regulation that limits the control of $F(c)$ in determining fee. The maximum fee allowed is 15 percent of target cost for research and development contracts. The form of fee as a function of c is most often expressed as follows:

$$F(c) = mc_0 + b \text{ for } 0 \leq c \leq c_0,$$
$$F(c) = mc + b \text{ for } c_0 \leq c \leq c_1,$$
$$F(c) = mc + b \text{ for } c \geq c_1,$$

where m and b are as defined under FPIF, c_0 is the cost corresponding to maximum fee, and c_1 is cost corresponding to minimum fee. In most cases b is determined from target cost c_T and target fee $F(c_T)$ which are specified by negotiation.

Cost Plus Fixed Fee Contract (CPFF)

In these contracts the cost is paid by the customer and in addition a fee is paid to the contractor. That fee which may be denoted by F_0 is independent of the cost.

A brief discussion of the relation between the sharing ratio and the slope of the share line will facilitate discussion of Table 1. Let x/y denote the sharing ratio where the contractor assumes y percent of any change in cost (as a loss in fee) and the customer assumes x percent (clearly $x + y = 100$).

The incentive fee share line has slope $m = \Delta F/\Delta c$ where Δc is positive and ΔF is negative resulting in a negative value for m. If cost increases an amount Δc the contractor assumes y percent of that increase as a loss in fee. In such a case $\Delta F = -y\Delta c/100$. Thus $m = -y/100$. Since the maximum share the contractor can assume is 100 percent the steepest share line slope is 45 degrees, corresponding to $m = -1$.

In the fee formulas used previously a simple linear form: $F = mc + b$, has been used for ease in presentation. It is more common to express the share line in terms of target cost, c_T, and target fee, $F_T = F(c_T)$. The resulting expression is the fee formula given by Eq. (2) which may be employed in ıny of the incentive contract forms above.

$$F(c) = m(c - c_T) + F_T. \tag{2}$$

The fee expressed by Eq. (2) clearly shows the role of the slope m and the share ratio x/y as developed above. Since m is negative any cost c below target will result in an increase in fee.*

Reference to the quantities R in Table 1 and employing the fact that $m^2 \leq 1$ established above results in the following:

R is greatest for FFP contracts and is zero for CPFF contracts in complete agreement with the usual risk levels attributed to such contracts.

For FPI and CPIF contracts R falls between the extreme values Var (C) and 0, as it should. It is a mathematical possibility for R(FPIF) to be less than R(CPIF) which would be counter to the desired relation. In practice, however, such a result would be based on extremely poor contract formulation.

---

*It is common in incentive contracting literature to write Eq. (2) in the form $F = m(c_T - c) + F_T$, where m has a positive value equal to the absolute value of the slope. Though simpler to express in some applications such a formula is rather difficult to understand in common mathematical terms such as are used in describing Eq. (2).

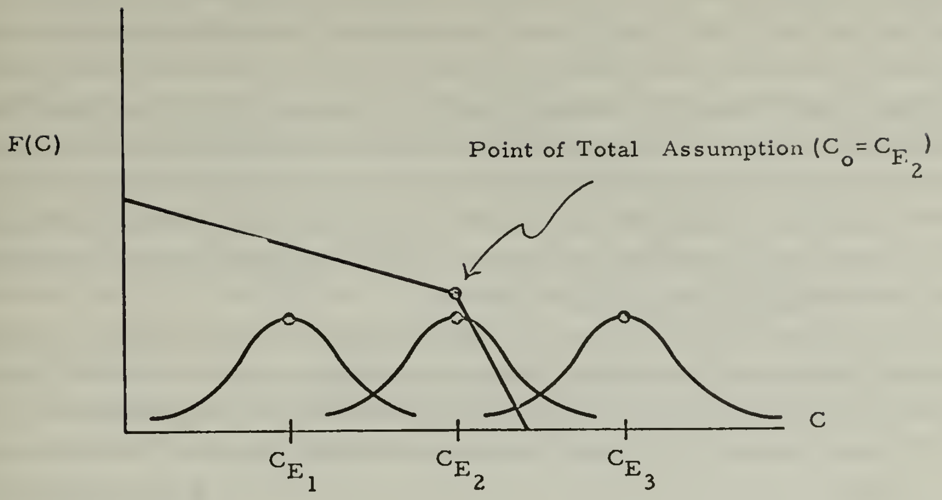


Figure 2. Dependence of R on expected cost C_E for a FPIF Contract.

The dependence of R on C_E is illustrated in Figure 2 for an FPI contract. Three distributions are used for the random variable cost. These distributions differ only in their expected value, all have the same variance σ^2. The expected values are denoted C_{E_1}, C_{E_2}, and C_{E_3}, respectively. For the distributions shown, it is assumed that the integrals used to define R are well approximated in each case by the following values on either side of the point of total assumption C_0:

$$R_1 = \text{Risk for } C_{E_1} \approx m^2\sigma^2 + 1(0) = m^2\sigma^2,$$

$$R_2 = \text{Risk for } C_{E_2} \approx m^2\left(\frac{\sigma^2}{2}\right) + 1\left(\frac{\sigma^2}{2}\right) = \frac{(m^2+1)}{2}\sigma^2,$$

$$R_3 = \text{Risk for } C_{E_3} \approx m^2(0) + 1(\sigma^2) = \sigma^2.$$

Since $m^2 < 1$ it follows that $R_1 < R_2 < R_3$ and in fact the R_3 corresponds to the same risk as in a FFP contract.

It is not meaningful to give detailed comparison between the different R values from Table 1 since they would normally be employed in different real world situations. This is expressed in terms of the uncertainty measure $\sigma^2 = \text{Var }(C)$ by observing that FFP contracts are used when σ^2 is small and CPFF contracts are used when σ^2 is very large with the other situations occupying intermediate positions. In terms of risk and the measure of risk, R, this states that the contractor assumes maximum risk in those situations where total uncertainty (risk) is least.* This is in agreement with the basic intent of having different types of contracts.

For CPIF contracts the form of R given in Table 1 allows some clarification of the role of range of incentive effectiveness (RIE), share line slope m, and the cost values c_0 and c_1. At a cost of c_0 the contractor earns maximum fee, and at c_1 he earns minimum fee. There is no motivation within the contract fee structure to drive cost below c_0 or, having reached c_1, to restrict increases in cost. The RIE is equal to $c_1 - c_0$; however, various values of c_0 and c_1 may be used to get the same RIE. The form

*This observation is well known and occurs in various forms throughout the contracting literature. For example, see Durbin, E. P., "Contingent Pricing Policies," Stanford Univ. Tech. Rept. 76, Contract Nonr 225(53) (Dec. 1964), p. 2, on the low total risk in FPI contracts.

of R from Table 1 shows that risk may be charged by variation in any of the quantities m, c_0, or c_1. The dependence of R on c_E will be discussed in the applications section by relating c_E to target cost c_T. The mathematical definition of R given by Eq. (1) requires c_E in the formulas of Table 1. In the real world c_E need not equal c_T though it is common in incentive contracting to act as though these two are the same. These observations on the form of R will be concluded by noting that in the CPIF case removal of maximum and minimum fee restrictions would result in an R value directly proportional to the R value for FFP contracts. The factor of proportionality is $m^2 < 1$, and as m tends to one, such contracts tend to become FFP. Thus, in general, increasing the RIE or increasing the slope of CPIF contracts increases the level of risk assumption in such contracts.

Further illustration of the risk measure R specified in Table 1 and of its variation for different contract types will be given in terms of numerical values for specific cases. The relation of R to real world uncertainty will be obtained by employing several different probability distributions for C. In these calculations it will be necessary to specify parameters of contract structure. Simple but reasonable values will be used for purpose of illustration. No actual contract values will be used since the purpose is to illustrate the concept of R rather than study specific real situations. Results will be presented in tabular form for ease in visual appreciation. The cases to be presented are described below.

As has been previously mentioned, different contract types would be used for different real world levels of uncertainty; however, it is useful for illustrative purposes to consider R for each contract type for each of several probability distributions on C. Three normal distributions will be used, each having $C_E = 20$ and with $\sigma^2 = 0.1$, 2.0, and 10.0, respectively. In addition a uniform distribution of the form $g_C(c) = 1/(25-15)$ will be used. This results in four different expressions of real world uncertainty. The measure R will be computed for each contract type in each real world model situation. For this purpose the following contract structure parameters will be used:

FPIF cases: The PTA value $c_0 = 25$ and the slope of the share line is $m = -0.5$ corresponding to a sharing ratio of 50/50.

CPIF cases: Location of maximum fee is $C_0 = 8$, location of minimum fee is $C_1 = 25$, and the slope of the share line is $m = -0.15$ corresponding to a sharing ratio of 85/15.

It may be remarked that the specific form for fee is not required in the calculation for R. In the applications section some discussion will be presented relating actual fee values to R in a meaningful way.

The 16 illustrative values for R are shown in Table 2, the expected cost value c_E is 20 in every case.

TABLE 2. Illustrative Values of R

Contract Type

Distribution of C	CPFF	CPIF	FPI	FFP
Normal $\sigma^2 = 0.1$	0	0.0023	0.025	0.1
Normal $\sigma^2 = 2.0$	0	0.046	0.504	2
Normal $\sigma^2 = 10.0$	0	0.176	4.29	10
Uniform $\sigma^2 = 8.33$	0	0.19	2.08	8.33

The values shown for R in Table 2 indicate the kinds of numbers obtained in typical cases. They satisfy the properties desired for R as discussed above. When $\sigma^2=0.1$ there is essentially no uncertainty in the real world situation, the risk index is low for all contract types. In every case risk is greatest for the FFP and the FPI risk is greater than the CPIF risk. Results in the uniform distribution case are interesting in comparison to the normal case having greater variance ($\sigma^2=10$). For lower risk assumption contracts the variability present in the uniform case produces a higher risk index than does the high variance normal. As more risk is assumed by the contractor the measure of risk becomes less sensitive to the spread in values produced by the uniform distribution. The measure of risk in those cases becomes more like the high variance normal case. Of course all the values in Table 2 depend on the contract parameters assumed and values of R can be produced having almost any form; however, the cases in Table 2 are felt to be typical and, as such, indicate realistic values for R. It may be noted that in each uncertainty case the contract type values differ from each other by an order of magnitude. This nicely expresses the normal attribution of differing levels of risk from one type contract to another.

APPLICATIONS OF *R*

This section discusses three applications of the definition for R given by Eq. (1) utilizing the forms exhibited in Table 1. A separate subsection will be devoted to each application.

Relation between target cost and expected cost

Various studies into the value of incentive contracts by Logistics Management Institute (LMI) [5], Fisher [2], Jones [3], and others, have pointed out the problems associated with improper selection of target cost c_T. These works combine statistical analysis, case studies, and economic augments in considering the effects of high or low target cost. The conclusions are often critical of incentive contracts on the basis of improper target cost. Such conclusions are responded to by others who point out that in any contract the parameters such as target cost should be properly selected. It is not intended to enter upon this debate here, but rather to clarify the distinction between C_T and the expected cost C_E. When researchers point out that C_T is improperly selected it may be presumed that they mean it is not equal to C_E. The effect this has on incentive fee payments is the subject of the references introduced above and will not be discussed here. The measure R, however, may be used to discuss the effect of improper selection of C_T on the presumed assumption of risk by the contractor for which it is being paid an actuarial fee. Though distinct from the researches mentioned above it is complementary to them and is, moreover, directly related to specific contract type by means of the differing forms assumed by R.

It will suffice for this discussion to study the relation given in Table 1 for R in the CPIF case. Consider the related expression given in Eq. (3):

$$H(A)=\int_{C_0}^{C_1}(c-A)^2 g_C(c)\,dc, \tag{3}$$

It is reasonable to assume that the probability of C falling below C_0 or above C_1 is small in an ideal CPIF contract.* On that assumption $H(A)$ is well approximated by $H^*(A)$ given in Eq. (4):

$$H^*(A)=\int_{-\infty}^{\infty}(c-A)^2 g_C(c)\,dc, \tag{4}$$

*In practice it is fairly common for $C>C_1$ due to difficulties in selecting and negotiating appropriate parameter values for contract specification; however, in theory a CPIF contract should have the low probability feature assumed here.

it is well known that $H^*(A)$ assumes its minimum value when $A = C_E$. Thus the effect on risk caused by $c_T \neq c_E$ can be simply described. When this occurs the measure of contractor risk R, based on c_T is greater than it would be if the assumed value of C_E (i.e., C_T) was replaced by the true C_E value. However, by acting like $c_E = c^T$ and using that value to compute R it may seem that the larger value R applied and the contractor might well expect a larger actuarial fee. Thus an improper target cost can result in increases in actuarial fee based on apparent levels of risk assumption that may not truly reflect an actual degree of real world uncertainty. This interpretation in terms of R presents an analytic expression for the widely held view that incorrect target cost can be used as a hedge against true risk assumption by a contractor. It might be remarked that this effect is quite distinct from the role of incorrect target cost in producing inflated incentive fees.

Selection of Contract Type

Different types of contracts have been developed to accommodate various situations particularly regarding the relative assumption of risk by the contractor. By using the measure R calculated for cases of interest (as illustrated in Tables 1 and 2), a quantification of the contractor's risk assumption may be obtained. From a knowledge of the environment under which the contract must be carried out, a decision can be made as to what contract types would be most appropriate. Such analyses would be useful to the customer when calling for proposals or in evaluating proposals for contractor selection. Contractors could also apply the quantification to present their position regarding contract type either when submitting a proposal or as part of negotiation procedures.

In using R, as described above, it would be particularly helpful if the scale of R values was expressed in units of the variance of C. Such a scale would always run from zero to one. As a guide to contract type selection that scale might be divided into several regions, each region corresponding to major contract types. An example of such a scale is shown in Figure 3, where R^* denotes $R/\mathrm{Var}(C)$.

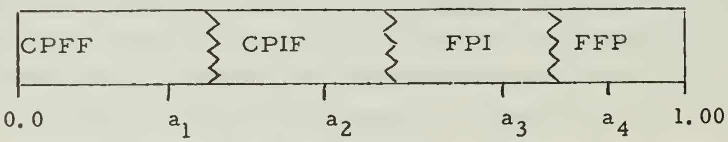


Figure 3. Contract type scale for risk.

The scale values a_i (illustrated for $i = 1, 2, 3, 4$) would be selected on the basis of experience and judgment. Such scales could be used to gain insight into the contract type selection process. It may be remarked that the scale would not be expected to be linear; low values would prevail rather far along till the FFP region was reached. This is true for the examples in Table 2 and reflects the conservative position of both contractor and customer regarding contractor risk assumption in those situations where non-FFP contracts are used (e.g., high cost, special situation Government contracts).

The measure R may be useful in understanding the problem of selecting the actuarial fee portion of target fee. Research into the behavior of R in various contract situations would provide guidelines as to the form the dependence between actuarial fee and R should take.

Research on Utility Functions

The utility function concept is widely used in theoretical studies of the decision process. It is a well accepted quantity in economic theory; however, specific details of a contractor's utility function are generally difficult or impossible to formulate. It would be valuable in many studies to have methods for achieving explicit quantification of a utility function (even if only in the form of a numerical table). Moreover, relations between the utility function and other quantities of economic theory are highly desirable. Equation (1) provides one such relation connecting the utility function to risk. By means of that relation one may carry out at least two kinds of theoretical studies on the structure of utility functions.

For a fixed contract structure and level of real world uncertainty (represented by $g_C(c)$), one may study the problem of selecting that class of utility functions that minimize R. This takes the mathematical form of a problem in the calculus of variations. By optimizing R over a prescribed class of utility functions (such as monotonic increasing functions) various restricted problems can be formulated. Such problems are difficult to solve, in general, and numerical solutions would normally be required. The utility function resulting would indicate how contractor value systems (as described by their utility functions) relate to contractor assumption of risk, other conditions being held fixed.

Another method for employing R to obtain forms for the utility function is to suppose that the index of risk for one contract type should be proportional to the index for another contract type. In such a case, the expression of uncertainty, $g_C(c)$, would be different for each index. Thus one might suppose that in a particular situation denoted by subscript 1, R for a CPIF contract should be half the value of R in another situation denoted by subscript 2, where a FPIF contract was appropriate. Such an assumption would lead to the functional relation given by Eq. (5) for the utility function $u(\cdot)$ which is common to both risk situations.

$$(5) \qquad \int_{c_1}[u(F_1(c))-\bar{u}_1]^2 dG_{C_1}(c)=\frac{1}{2}\int_{C_2}[u(F_2(c))-\bar{u}_2]^2 dG_{C_2}(c),$$

where

$$\bar{u}_i=\int_{c_i} u(F_i(c))dG_{c_i}(c),\ i=1,2.$$

CONCLUSIONS

The measure of contractor risk assumption introduced by Eq. (1) seems to provide a useful quantification of risk. It is an unambiguous quantity which is specifically related to individual contract forms, contractor utility for money, and the level of uncertainty prevailing in nature. Use can be made of that index in several studies dealing with contract theory and the study of utility functions. Though there is a wide literature on risk theory as studied for example by Pratt [6] and by Lipson [4], explicitly defined expressions seem to be needed which will allow quantitative analysis of different risk situations. The present paper has shown the theoretical advantages of R.

REFERENCES

[1] Durbin, E. P., "Contingent Pricing Policies," Stanford Univ. Tech. Rept. 76, Contract Nonr 225(53) (Dec. 1964).
——— "Pricing Policies Contingent on Observed Product Quality," Technometrics, Vol. 8, No. 1, (Feb. 1966).

[2] Fisher, Irving, "Cost Incentive and Contract Outcomes: An Empirical Analysis," Rand Corp. (Sept. 1966).
——— "Improving the Effectiveness of Incentive Contracting," Rand Paper P-3870 (July 1968).

[3] Jones, Troy H., "A Study of Selected Aspects of the Use of Incentive Contracts," Ph.D. dissertation, Ohio State Univ. (1967).

[4] Lipson, Melvin W., "Formalizing Risk in Incentive Contracting," Nat. Contract Manag. Assoc. Jour., Vol. 1, No. 4 (Fall 1967).

[5] Logistics Management Institute, "An Examination of the Foundations of Incentive Contracting," Report for Task 66-7 (May 1968).

[6] Pratt, John W., "Risk Aversion in the Small and in the Large," Econometrica, Vol. 32, No. 1-2 (Jan.-Apr. 1964).

* * *

APPLICATIONS OF A GENERALIZED COMBINATORIAL PROBLEM OF SMIRNOV

Ronald Alter
University of Kentucky

and

B. P. Lientz
System Development Corporation

ABSTRACT

In this paper applications of results obtained by these authors for a generalized version of a problem proposed by Smirnov, are considered. The areas of application explored are system interface, queueing, transportation flow, and sequential analysis. The included table should be invaluable to the reader in applying these results. Finally the relationship between the limiting and exact expressions relating to this table is also explored.

1. INTRODUCTION

In 1966, N. V. Smirnov proposed the following combinatorial problem: given n objects of $s+1$ classes, how many ways can the n objects be arranged in a chain so that adjacent objects belong to separate classes? In Ref. [2], Sarmanov and Zaharov interpreted the problem in terms of transitions from class to class. They applied a result from the theory of Markov chains and got an exact expression for the number of chains. In this exact case, however, their formula is based upon knowledge of all possible transitions from class to class, which is tantamount to knowledge of the number of possible chains. The main results in Ref. [2] are limiting formulas, obtained from the exact expressions, for the two special cases of $s=2$ and the case in which the number of each class is the same.

For the limiting case of $s=2$, Sarmanov and Zaharov obtained

$$M^{(3)}(r_1, r_2, r_3)=M_1^{(3)}(r_1, r_2, r_3)+M_2^{(3)}(r_1, r_2, r_3), \tag{1.1}$$

$$M_1^{(3)}(r_1, r_2, r_3)=nF(r_1, r_2, r_3)(1+o(1)), \tag{1.2}$$

$$\begin{aligned} &M_2^{(3)}(r_1, r_2, r_3) \\ &=\{r_1F(r_1-1, r_2, r_3)+r_2F(r_1, r_2-1, r_3)+r_3F(r_1, r_2, r_3-1)\}(1+o(1)), \end{aligned} \tag{1.3}$$

and

$$\begin{aligned} &F(r_1, r_2, r_3) \\ &=\frac{\sqrt{2\pi}\, r_1!r_2!r_3![(n^2-2r_1^2-2r_3^2-2r_2^2)(n-2r_1)(n-2r_2)(n-2r_3)]^{1/2}}{8r_1r_2r_3\Gamma^2(1+n/2-r_1)\Gamma^2(1+n/2-r_1)\Gamma^2(1+n/2-r_2)\Gamma^2(1+n/2-r_3)} \end{aligned} \tag{1.4}$$

r_I is the number of objects in the Ith class. $M_1^{(s+1)}(\ldots)$ is the number of chains with objects of different classes at the ends, $M_2^{(s+1)}(\ldots)$ is the number of chains with objects of the same class at the ends. $M^{(s+1)}(\ldots)$ is the total number of chains. In (1.4) $\Gamma(.)$ is the gamma function.

For the limiting case when the number in each class is the same $(r_1=r_2=\ldots=r_{s+1}=n/(s+1)=r)$ they showed that

$$M^{(s+1)}(r, \ldots, r)=M_1^{(s+1)}(r, \ldots, r)+M_2^{(s+1)}(r, \ldots, r), \tag{1.5}$$

where

$$M_1^{(s+1)}(r, \ldots, r) = (2\pi r)^{-s/2}\left(\frac{s-1}{s+1}\right)^{s/2}\sqrt{s+1}\,s^n(1+o(1)) \tag{1.6}$$

and

$$M_2^{(s+1)}(r, \ldots, r) = \frac{1}{s}M_1^{(s+1)}(r, \ldots, r)(1+o(1)). \tag{1.7}$$

Because the two cases considered in Ref. [2] are quite restrictive it might be hoped that further limiting results could be obtained by extending the arguments in Ref. [2]. This is not easily done because the method in Ref. [2] is based upon finding the coordinates of a stationary point for which the expression in (1.2) is maximized. To do this one must consider s^2-s-1 nonlinear equations, except in the case of $s=2$. Using this approach, simplication for $s>2$ is possible only for the case where the r_I's are all equal.

In Ref. [1] the authors devote their attention to the finite (or exact) case. Workable representation theorems for $M^{(s+1)}(\ldots)$ are attained; however, for the objectives of the present paper we only need Theorem 1 of Ref. [1], namely:

$$M^{(s+1)}(r_1, \ldots, r_{s+1}) = \sum_{I=1}^{s+1}\sum_{J=1}^{r_I}(-1)^{J+1}M^{(s+1)}(r_1, \ldots, r_{I-1}, r_I-J, r_{I+1}, \ldots, r_{s+1}), \tag{1.8}$$

where

$$M^{(s+1)}(r_1, \ldots, r_{I-1}, 0, r_{I+1}, \ldots, r_{s+1}) = M^{(s)}(r_1, \ldots, r_{I-1}, r_{I+1}, \ldots, r_{s+1}). \tag{1.9}$$

These formulas give us a rapid method of getting $M^{(s+1)}(\ldots)$. The expressions also indicate that the quantities $M^{(s+1)}(\ldots)$ can be obtained from those of $M^{(s)}(\ldots)$, recursively, as follows. If we wish to compute $M^{(s+1)}(r_1, r_2, r_3, r_4)$, we first decompose this by (1.8). For cases in which any r_I-J is zero, we use (1.9). For the case when any r_I-J is one, we can obtain $M^{(s+1)}(r_1, \ldots, r_{I-1}, 1, r_{I+1}, \ldots, r_{s+1})$ directly from

$$\left(\sum_{K\neq I} r_K\right)M^{(s)}(r_1, \ldots, r_{I-1}, r_{I+1}, \ldots, r_{s+1}).$$

For the case when all arguments in $M^{(s+1)}(\ldots)$ are greater than one, we can successively apply (1.8) to obtain a further decomposition. This is a workable procedure which, if we know the values of $M^{(3)}(r_1, r_2, r_3)$ for various values of r_1, r_2, and r_3, we can use to compute $M^{(s+1)}(r_1, \ldots, r_{s+1})$ for any s exceeding two. This is valuable in light of the knowledge of the asymptotic results for $s=2$.

In Section 2 specific applications of the above results to problems in system interface, transportation flow and sequential analysis are examined. The discussion in the preceding paragraph indicates the importance of the case $s=2$ and how it can be used to obtain $M^{(s+1)}(\ldots)$ for larger values of s. Section 3 contains a description of the table of values of $M^{(3)}(\ldots)$ which appears in Section 4, and explores the relationship between the limiting and exact expression.

2. APPLICATIONS

Several applications to specific areas of operations research and statistics will be examined in this section.

The first application is concerned with a problem encountered in system interface. The situation can be described as follows. A number of sources input messages into a central data center. These messages are stored at the center and from time-to-time are transmitted back to the sources and/or to

other receivers. The messages themselves contain information about some collection of objects or subjects (e.g., identification codes, heading, altitude, and so forth.). The messages from all sources that refer to a given object are stored in the same location in the data center. Message overlay occurs when information about a particular object obtained from one source is eradicated or altered by a later message from a different source. The importance of overlay can be seen in situations where the receivers must implement a course of action (such as deploying forces) which is dependent upon this information. For example, messages which differ with respect to altitude or heading could be transmitted to a receiver. This would obviously have an impact on the interception strategy.

One important part of the problem is the computation of the number of ways message overlay can occur. This would give an idea of what corrective action could be taken so as to minimize overlay. For example, a possible corrective step would be to establish pre-storage comparison and have notification of the existence of overlay given to the sources of the messages in the conflict. Another step would be to increase storage capacity and employ a selective retrieval system. In the notation above $s+1$ is the number of sources and r_l refers to the number of messages sent on any particular object by the lth source to the central data center. The exact expressions given in (1.8) and (1.9) can be used to compute the number of ways message overlay can occur. In one problem encountered thus far, there were intermediate receivers between the initial sources and the data center. The purpose of these receivers was to carry out some degree of preprocessing and sorting. The intermediaries were in effect sources to the data center. The effect of this was to reduce the number of sources.

A second application is one dealing with a queueing situation. We suppose that there is a warehouse or comparable storage facility which dispenses spare parts. The customers for the warehouse usually require parts which can be grouped into several distinct categories (e.g., types of transmissions, engine blocks, and the like). Within each category the parts are assumed to be sufficiently similar so as to be located in the same area of the warehouse. Each server is initially assumed to be capable of locating and dispensing all parts in the warehouse. Service would perhaps be improved if servers were assigned to particular categories. In the notation of the previous section, $s+1$ is the number of categories. The number of ways different categories can be called in succession can be computed. The planner can use this to evaluate the possibility of allocating servers to particular categories. For example, suppose a sample is obtained for the observed number of different successive categories and found to be close to the maximal number determined by the previous section. He then knows that if the input demands remain relatively stable, he will probably not gain appreciably by allocating servers to categories.

Another application is in the study of traffic flow. Engineers are studying the feasibility of assigning specific lanes to types of traffic (e.g., only trucks and buses permitted in the far right-hand lane). If the traffic in the affected lanes is already composed almost entirely of the desired vehicle type there is no need for assignation. If alternately there are chains of vehicles of different types, then action might be necessary. The formulas in Section 1 compute the maximal mixture for varying classes of vehicles. This would be helpful in determining the threshold beyond which corrective action is necessary.

In order statistics and sequential analysis there is sometimes the problem of having sample points from various populations intermingling. To determine the likelihood of various occurrences the maximum mixture of observations from various populations is needed.

3. CONSTRUCTION AND DESCRIPTION OF TABLE

The table was constructed with the use of Egs. (1.8) and (1.9) in Section 1. An IBM 360 computer

was used with Fortran IV language. The computation of the quantity $M^{(3)}$ (. . .) was carried out in fixed point mode and the calculation of the combinations was carried out in floating point mode.

For convenience, the table has five columns. The first column gives the total number of objects. n, The second, third, and fourth columns give the numbers of objects in the first, second, and third classes, respectively. The last column gives the number of allowable chains (i.e., chains in which no two objects of the same class are adjacent). It should be noted that the table includes values of $M^{(3)}$ (. . .) only for combinations which lead to nonzero $M^{(3)}$ (· . . .) and for $r_1 \leq r_2 \leq r_3$.

The table was carried out to $n=20$. Computations were done for higher orders of n. However, the limiting formulas given by (1.1)-(1.4) were found to reasonably approximate the exact expressions for $n > 20$.

4. TABLE OF NUMBER OF CHAINS FOR THREE CLASSES

n	r_1	r_2	r_3	$M^{(3)}(r_1, r_2, r_3)$	n	r_1	r_2	r_3	$M^{(3)}(r_1, r_2, r_3)$
3	1	1	1	6	16	4	6	6	7580
4	1	1	2	4	16	4	5	7	3930
5	1	2	2	12	16	3	6	7	1990
5	1	1	3	2	16	2	7	7	450
6	2	2	2	30	16	4	4	8	420
6	1	2	3	10	16	3	5	8	322
7	2	2	3	38	16	2	6	8	140
7	1	3	3	18	16	1	7	8	30
7	1	2	4	3	17	5	6	6	22274
8	2	3	3	74	17	5	5	7	13140
8	2	2	4	24	17	4	6	7	9303
8	1	3	4	14	17	3	7	7	3118
9	3	3	3	174	17	4	5	8	2331
9	2	3	4	79	17	3	6	8	1323
9	1	4	4	24	17	2	7	8	393
9	2	2	5	6	17	1	8	8	48
9	1	3	5	4	17	4	4	9	70
10	3	3	4	248	17	3	5	9	56
10	2	4	4	138	17	2	6	8	28
10	2	3	5	44	17	1	7	9	8
10	1	4	5	18	18	6	6	6	48852
11	3	4	4	480	18	5	6	7	31632
11	3	3	5	212	18	4	7	7	15324
11	2	4	5	135	18	5	5	8	10500
11	1	5	5	30	18	4	6	8	7854
11	2	3	6	10	18	3	7	8	3164
11	1	4	6	5	18	2	8	8	594
12	4	4	4	1092	18	4	5	9	812
12	3	4	5	588	18	3	6	9	504
12	2	5	5	222	18	2	7	9	184
12	3	3	6	100	18	1	8	9	34
12	2	4	6	70	19	6	6	7	80672
12	1	5	6	22	19	5	7	7	56040
13	4	4	5	1668	19	5	6	8	32634
13	3	5	5	1026	19	4	7	8	17724
13	3	4	6	445	19	3	8	8	4804
13	2	5	6	206	19	5	5	9	5572
13	1	6	6	36	19	4	6	9	4354
13	3	3	7	20	19	3	7	9	2024
13	2	4	7	15	19	2	8	9	509
13	1	5	7	6	19	1	9	9	54
14	4	5	5	3228	19	4	5	10	126
14	4	4	6	1700	19	3	6	10	84
14	3	5	6	1150	19	2	7	10	36
14	2	6	6	326	19	1	8	10	9
14	3	4	7	190	20	6	7	7	156664
14	2	5	7	102	20	6	6	8	100044
14	1	6	7	26	20	5	7	8	73416
15	5	5	5	7188	20	4	8	8	27888
15	4	5	6	4315	20	5	6	9	23828
15	3	6	6	1882	20	4	7	9	14168
15	4	4	7	1110	20	3	8	9	4728
15	3	5	7	806	20	2	9	9	758
15	2	6	7	292	20	5	5	10	1764
15	1	7	7	42	20	4	6	10	1428
15	3	4	8	35	20	3	7	10	744
15	2	5	8	21	20	2	8	10	234
15	1	6	8	7	20	1	9	10	38
16	5	5	6	11492					

REFERENCES

[1] Alter, Ronald and B. P. Lientz, "A Generalization of a Combinatorial Problem of Smirnov," System Development Corporation Document, SP 3254.

[2] Sarmanov, O. V. and V. K. Zaharov, "A Combinatorial Problem of Smirnov." *Dokl. Akad. Nauk. SSSR. 176*, 1147–1150 (1967).

SINGLE MACHINE SEQUENCING WITH RANDOM PROCESSING TIMES AND RANDOM DUE-DATES*

Thomas B. Crabill
University of North Carolina

and

William L. Maxwell
Cornell University

ABSTRACT

A single machine scheduling problem in which both the processing times and due-dates of the jobs awaiting servicing are random variables is analyzed. It is proved that the properties of the shortest processing time rule and the due-date rule which are known for the deterministic situation also hold in the probabilistic environment when they are suitably, and reasonably, refined for this context.

Several authors [1–7] have considered the problem of specifying an optimal sequence in which to schedule a fixed set of jobs at a single machine. Most of this work assumes that the processing times, due-dates (desired completion times), and penalty costs are known with certainty. Two exceptions are the works of Banerjee [1] and Rothkopf [5], in both the processing times are random variables; with one exception, however, these papers do not relate directly to the properties of scheduling procedures which are known to be optimal in a completely deterministic environment. In this paper we show that with a variety of objective functions the obvious counterparts of optimal rules in the deterministic situation are also optimal in the probabilistic context.

The operational model is a situation in which there are n jobs waiting for service before a single machine. The machine is ready to start processing the jobs and, for convenience, we assume that a calendar time of zero corresponds to the commencement of service on the first job selected. We also assume that once the processing of a job is initiated, it is processed to completion. The scheduling decision is, then, to decide upon which one of the n! possible ordering of jobs to specify *at time 0.*†

The notation used is:

P_i = a random variable representing the processing time of job i, $i=1, 2, \ldots, n$;

$G_P(\cdot)$ = the joint distribution function of the processing times, $P=(P_1, P_2, \ldots, P_n)$;

$G_{P_i}(\cdot)$ = the marginal distribution function of the processing time, P_i;

D_i = a random variable representing the due-date of job i, $i=1, 2, \ldots, n$;

$H_D(\cdot)$ = the joint distribution function of the due-dates, $D=(D_1, D_2, \ldots, D_n)$;

$H_{D_i}(\cdot)$ = the marginal distribution function of the due-date, D_i;

u_i = a weighting coefficient, or value, for job i, $i=1, 2, \ldots, n$;

$[i]$ = the job scheduled to be processed in the ith position in sequence. For example, if jobs 1, 2, 3 are scheduled in the order 2, 1, 3, then $[1]=2$, $[2]=1$, and $[3]=3$.

---

*This work was partially supported by National Science Foundation contract NSF–GP–2729.

†It is important to note that rescheduling is not allowed; there can be no subsequent rearrangement of the sequence in which jobs are performed based upon the realized completion times of a set of jobs.

The commonly used measures of performance are related to the flow time, lateness, and tardiness of jobs [2]. The flow time, F_i, of a job is

$$F_{[i]} = \sum_{j=1}^{i} P_{[j]};$$

the lateness, L_i, of a job is

$$L_i = F_i - D_i;$$

and the tardiness, T_i, of a job is

$$T_i = \max(L_i, 0).$$

When both processing times and due-dates are deterministic and known, the following results have been proved:

1. The average flow time, $(1/n)\Sigma F_i$, and the average lateness, $(1/n)\Sigma L_i$, are minimized by scheduling the jobs in order of increasing processing time [7] (called in [2], SPT, or the shortest processing time rule). The scheduling of jobs in the reverse order of processing time results in the maximum flow time [2]. These results are also obviously true when due-dates are stochastic in nature and expected values of average flow times and average lateness are considered.

2. The maximum job lateness and maximum job tardiness are minimized by scheduling the jobs in order of increasing due-dates [3] (called in [2] DDATE, or the due-date rule).

3. The minimum job lateness is maximized by scheduling the jobs in the order of the difference between due-date and processing time [2] (called in [2] SLACK, or the slack rule).

4. Weighted average flow time, $(1/n)\Sigma u_i F_i$, and weighted average lateness, $(1/n)\Sigma u_i L_i$, are minimized by scheduling the jobs in order of increasing ratio of P_i/u_i [7]. Case 1 is obviously a special example of Case 4 with all u_i's equal. As in Case 1, these results obviously obtain when due-dates are random variables and expected values are considered.

When the processing times are random variables and the due-dates are deterministic:

5. Rothkopf [5] has shown that the expected weighted average flow time weighted $E[(1/n)\Sigma u_i F_i]$, is minimized by scheduling the jobs in order of increasing ratio of $E[P_i]/u_i$. It follows that the expected weighted average lateness, $E[(1/n)\Sigma u_i L_i]$ is also minimized by this order.

6. Banerjee [1] considered the objective of minimizing,

$$\max_i \left\{ \mathrm{Prob}(L_i \geqq 0) \right\},$$

and showed that this is achieved by scheduling the jobs in due-date order.

This paper extends the above conditions to the case in which both processing times and due-dates are random variables. We have noted that the rules applicable for weighted mean flow time and weighted mean lateness trivially extend to this environment by scheduling according to the ratio $E[P_i]/u_i$. The fact that DDATE is the appropriate schedule to use when measured against the maximum lateness is not obvious, and the fact that there are several certainly equivalent interpretations of this random variable make the problem interesting. Surprisingly, scheduling according to due-date proves to be optimal for several plausible criteria.

MINIMIZING THE MAXIMUM PROBABILITY OF LATENESS

This is the objective function studied by Banerjee [1]. Theorem 1 extends Banerjee's results to a situation in which the due-dates are random variables. The concept of stochastic ordering is employed.

Two random variables, X and Y, are said to be stochastically ordered if their distribution functions, say $F_X(\cdot)$ and $F_Y(\cdot)$, satisfy $F_X(z) \geqq F_Y(z)$ for all z; when this is true, Y is said to be stochastically greater than X.

THEOREM 1: If the due-dates are stochastically ordered random variables, independent of the processing times, then minimizing

$$\max_i \text{Prob}(L_i \geqq 0)$$

is achieved by scheduling according to the sequence,

$$H_{D_{[1]}}(z) \geqq H_{D_{[2]}}(z) \geqq \ldots \geqq H_{D_{[n]}}(z), \text{ any } z.$$

Proof is by an interchange argument, a type of argument used for all of the previous results cited. Assume a schedule, S, in which two adjacent jobs, $[i]$ and $[i+1]$, are not in the order indicated, (i.e., $H_{D_{[i]}} < H_{D_{[i+1]}}$), and consider the effect of interchanging these jobs, producing a schedule S'. Note first that this action will affect only the probability of lateness of jobs $[i]$ and $[i+1]$ since the realizations of their processing times and due-dates are independent of the positions in which they are scheduled. Let M be the maximum probability of lateness of all other jobs in the schedule, except $[i]$ and $[i+1]$;

$$M = \max_{[j] \neq [i],\, [i+1]} \text{Prob}(L_{[j]} \geqq 0).$$

Under schedule S the value of the objective is

$$\max \{A, B, M\},$$

where

$$A = \text{Prob}\ (L_{[i]} \geqq 0) = \text{Prob}\ (P_{[1]} + P_{[2]} + \ldots + P_{[i-1]} + P_{[i]} \geqq D_{[i]})$$

and

$$B = \text{Prob}\ (L_{[i+1]} \geqq 0) = \text{Prob}\ (P_{[1]} + P_{[2]} + \ldots + P_{[i-1]} + P_{[i]} + P_{[i+1]} \geqq D_{[i+1]}),$$

and under schedule S' the value of the objective is

$$\max \{C, D, M\},$$

where

$$C = \text{Prob}\ (L_{[i]} \geqq 0) = \text{Prob}\ (P_{[1]} + P_{[2]} + \ldots + P_{[i-1]} + P_{[i]} + P_{[i+1]} \geqq D_{[i]})$$

and

$$D = \text{Prob}\ (L_{[i+1]} \geqq 0) = \text{Prob}\ (P_{[1]} + P_{[2]} + \ldots + P_{[i-1]} + P_{[i+1]} \geqq D_{[i+1]}).$$

We show that $\max \{A, B, M\} \geqq \max \{C, D, M\}$, thus demonstrating the desirability of interchanging jobs $[i]$ and $[i+1]'$ This requires two steps:

(a) $$B \geqq C$$

Let $\phi_{[i+1]}(\cdot)$ be the distribution function of

$$\sum_{j=1}^{i+1} P_{[i]}.$$

Since the processing times are independent of the due-date, we have

$$B=\int_{x=0}^{\infty}\int_{y=0}^{x} dH_{D_{[i+1]}}(y)\ d\phi_{[i+1]}(x)=\int_{x=0}^{\infty} H_{D_{[i+1]}}(x)\ d\phi_{[i+1]}(x)$$

and

$$C=\int_{x=0}^{\infty}\int_{y=0}^{x} dH_{D_{[i]}}(y)\ \ d\phi_{[i+1]}(x)=\int_{x=0}^{\infty} H_{D_{[i]}}(x)d\phi_{[i+1]}(x)$$

Since $H_{[i+1]}(x) \geqq H_{D_{[i]}}(x)$ for all x, if follows that $B \geqq C$.

(b) $$B \geqq D$$

$$\text{If } Q=\sum_{j=1}^{i-1} P_{[j]} + P_{[i+1]}, \text{ then,}$$

$$B=\text{Prob }(Q+P_{[i]} \geqq D_{[i+1]}) \geqq \text{Prob }(Q \geqq D_{[i+1]})=D.$$

follows immediately from the fact that $P_{[i]}$ is a positive random variable.

Since $B \geqq C$ and $B \geqq D$ imply max $(A, B, M) \geqq$ max (C, D, M), jobs $[i]$ and $[i+1]$ must be interchanged, resulting in schedule S'. Repeated application of this criterion results in an ordering which is equivalent to ordering the jobs by due-dates.

An example of stochastically ordered due-dates would be the system $D_i=D_{i+1}+R_i$, where R_i is a nonnegative random variable. This system, in fact, possesses a much stronger property which is used in Theorem 2; namely, Prob $[D_1 \leqq D_2 \leqq \ldots \leqq D_n]=1$.

MAXIMIZING THE PROBABILITY THAT EVERY JOB IS EARLY

The criterion is to maximize

$$\text{Prob }[L_1 \leqq 0, L_2 \leqq 0, \ldots, L_n \leqq 0].$$

THEOREM 2: If the joint due-date distribution is such that the jobs can be ordered so that Prob $[D_{[1]} \leqq D_{[2]} \leqq \ldots \leqq D_{[n]}]=1$ then this ordering maximizes the probability that every job is early.

Proof is, again, by an interchange argument. Assume a schedule, S, in which two adjacent jobs, $[i]$ and $[i+1]$, have Prob $[D_{[i+1]} \leqq D_{[i]}]=1$. Consider a schedule, S', in which jobs $[i]$ and $[i+1]$ are reversed in order and let C_S and $C_{S'}$ be the values of the objective function under S and S', respectively. Then,

$$C_S=\text{Prob }[M, A, B]$$

and

$$C_{S'}=\text{Prob }[M, C, D],$$

where M is the event that all other jobs are not late;

A is the event $Q+P_{[i]} \leqq D_{[i]}$;

B is the event $Q+P_{[i]}+P_{[i+1]} \leqq D_{[i+1]}$;

C is the event $Q+P_{[i]}+P_{[i+1]} \leqq D_{[i]}$;

D is the event $Q+P_{[i+1]} \leqq D_{[i+1]}$; and

$Q=\sum_{j=1}^{i-1} P_{[j]}$, and "," indicates the intersection of events.

Since Prob $[D_{[i+1]} \leqq D_{[i]}]=1$ and $P_{[i+1]}$ is a positive random variable event, B is a subset of the event A. Thus,

$$C_S = \text{Prob } [M, B, B].$$

Also, the event B is a subset of the event D, since $P_{[i]}$ is a positive random variable, so that,

$$C_S = \text{Prob } [M, B, D].$$

when $C_{S'}$ is written as the sum of the probabilities of two disjoint events

$$\begin{aligned} C_{S'} &= \text{Prob } [M, B, D] \\ &\quad + \text{Prob } [M, D_{[i+1]} < Q+P_{[i]}+P_{[i+1]} \leqq D_{[i]}, D], \\ C_{S'} &= C_S + \text{Prob}[M, D_{[i+1]} \leqq Q+P_{[i+1]} \leqq D_{[i]}, D]. \end{aligned}$$

The second term is clearly nonnegative, so that $C_{S'} \geqq C_S$ and thus schedule S' is preferable.

The conditions of Theorem 2 imply the conditions of Theorem 1. Thus the same job ordering will achieve both criteria if the stronger order conditions of Theorem 2 are satisfied (and the processing times are independent of due-dates).

MAXIMIZING THE PROBABILITY THAT THE MAXIMUM LATENESS IS MINIMAL

THEOREM 3: If the jobs can be ordered so that Prob $[D_{[1]} \leqq D_{[2]} \leqq \ldots \leqq D_{[n]}]=1$ then this ordering maximizes Prob $[\max_{[i]} L_{[i]} \leqq z]$ for all z.

The proof is by an interchange argument. By omitting the details,

$$\text{Prob } [\max_{[i]} L_{[i]} \leqq z] = \text{Prob } [P_{[1]} \leqq z+D_{[1]}, P_{[1]}+P_{[2]} \leqq z+D_{[2]}, \ldots].$$

By using the same subset argument as in the proof of Theorem 2, with $z+D_{[i]}$ replacing $D_{[i]}$ and $z+D_{[i+1]}$ replacing $D_{[i+1]}$, the proof follows.

COROLLARY: The expected value, $E[\max_{[i]} L_{[i]}]$ is minimized by this ordering.

This follows from the fact that $E(Y)=-\int_{-\infty}^{0} F(y)\,dy+\int_{0}^{\infty} (1-F(y))\,dy$ where $F(\cdot)$ is the distribution function of the random variable y.

COROLLARY: Prob $[\max_{[i]} T_{[i]} \leqq z]$ is maximized for all z and thus $E[\max_{[i]} T_{[i]}]$ is minimized by this ordering.

Since $T_{[i]}=\max[0, L_{[i]}]$, $\max_{[i]} T_{[i]}=\max[0, \max_{[i]} L_{[i]}]$,

Then

$$\text{Prob}[\max_{[i]} T_{[i]} \leqq z] = \begin{cases} 0 & z<0 \\ \text{Prob}[\max_{[i]} L_{[i]} \leqq z] & z \geqq 0. \end{cases}$$

By Theorem 3 the terms with $z \geqq 0$ are maximized by the ordering given in Theorem 3, for all i, and the result follows.

MINIMIZING THE MAXIMUM EXPECTED LATENESS

THEOREM 4: If the processing times and due-dates have any arbitrary joint distribution, then sequencing the jobs in the order

$$E(D_{[1]}) \leqq E(D_{[2]}) \leqq \ . \ . \ . \ \leqq E(D_{[n]})$$

minimizes $\max_{[i]} [E(L_{[i]})]$.

Noting that

$$E(L_{[i]}) = \sum_{j=1}^{i} E(P_{[j]}) - E(D_{[i]}),$$

the result is established in the same manner as the deterministic equivalent.

This criterion is not the same as $E[\max_{[i]} L_{[i]}]$, for which the stronger assumption of Theorem 3 is required. The conditions of Theorem 3 imply the above conditions, but the converse is not true.

REFERENCES

[1] Banerjee, B. P., "Single Facility Sequencing with Random Execution Times," Operations Research, Vol. 13, No. 3 (May 1965).
[2] Conway, R. W., W. L. Maxwell, and L. W. Miller, Theory of Scheduling (Addison-Wesley, Reading, Mass., 1967).
[3] Jackson, J. R., "Scheduling a Production Line to Minimize Maximum Tardiness," Management Sciences Research Project, UCLA (Jan. 1955).
[4] McNaughton, R., "Scheduling with Deadlines and Loss Functions," Management Science, Vol. 6, No. 1 (Oct. 1959).
[5] Rothkopf, M. H., "Scheduling with Random Service Times," Management Science, Vol. 12, No. 9 (May 1966).
[6] Schild, A. and I. J. Fredman, "Scheduling Tasks with Linear Loss Functions," Management Science, Vol. 7, No. 3 (Apr. 1961).
[7] Smith, W. E., "Various Optimizers for Single-Stage Production," Nav. Res. Log. Quart. **3**, 59–66 (1956).

* * *

ON THE OUTPUT OF PARALLEL EXPONENTIAL SERVICE CHANNELS

Robert R. Read
Naval Postgraduate School
Monterey, California

ABSTRACT

The paper treats the output process of a service center that has a large number of independent exponential channels in parallel. Initially all channels are working and there is a fixed backlog of items awaiting service. The moments are derived and central limit theorems are developed. Problems of computation are discussed and suitable formulae are developed. The joint distribution of the output of the center with the center's total busy time and total idle time are derived. Normal approximations to these distributions are presented.

INTRODUCTION

Consider a service center having c independent service channels. We are concerned with the number of customers served in time t given the original number of customers in the center and under the condition that no new customers enter the system in the time period $(0, t)$. Each channel will be assumed to have the same exponential service time distribution.

Such a center appears in many models. Consider, for example, the performance of a system having c identical units operating simultaneously and backed up by a bank of spares. Replacement from this bank takes place immediately after each failure and, when the bank is depleted, the system may be forced to operate at less than full strength. Such a system may fail to complete its mission if the decline in performance in $(0, t)$ is sufficiently great. (We note in passing that this situation, when cast as the repairman problem (see Ref. [1]), is the degenerate version of that problem when repair time is infinite.)

This process also appears in some life testing problems where the method of truncated testing is being used, see Epstein [5] and Epstein and Sobel [6]. In the testing of larger subsystems it commonly occurs that facilities are limited, i.e., there are only c test chambers and a fixed number of subsystems available for test. Testing takes place with replacement until all available subsystems are (or have been) on test, and then continues without replacement. Truncated test plans involve stopping either after fixed time or after a fixed number of failures. The reliability estimate is a function of the total test time accumulated by all subsystems, [5] and [6]. For purposes of experimental design and scheduling, the joint distribution of total test time and the output of the process is of interest. This is developed herein. Since total test time and total idle time (for the service center) add to a fixed quantity, we need only study one of them.

Collins and Guthrie [3] used this service center in modeling the carrier aircraft maintenance space problem (see Saaty [11], p. 181) as a discrete Markov chain. My own interest in the problem was developed when my student, LCDR G. M. Lanman [9], attempted to apply the Collin-Guthrie model to a larger problem. This study demonstrated the need for understanding this process and developing computational methods. The maintenance areas aboard an aircraft carrier served as the c service channels and that is the reason for the service center terminology used herein. Moreover, this

application stimulates the imagination to the extent that one can picture this service center as appearing in broad classes of larger models.

Of course, this center appears as part of the multiple server queue. Queueing problems require that the input process be specified. In the case of Poisson arrivals, Burke [2] and Reich [10] have shown that the output process matches the input process. The steady-state Markov chain has been characterized by Kendall [8] for general independent input. The transient case has been considered by Saaty [12] who developed some explicit formulae for $c=2$. This paper does not explicitly treat queueing systems; the input process is omitted throughout.

The paper contains the derivation of the transition probabilities, the first and second moments, and central limit theorems for this process. Problems of computation are discussed and suitable formulae are recorded. Then the joint distributions of the output of the process with the center total idle time (or total test time) are developed, both for fixed time and fixed number of failures. Also normal approximations are justified. Here the jargon is switched to that of life testing.

BASIC REPRESENTATION

Let X_t represent the number of customers in the service center at time t. The hypothesis that the center contains c independent identical exponential service channels implies that the output process $X(t)$ is Markovian, a result that effects substantial simplification. Let

$$P_{ij}(t)=Pr\{X_t=j|X_o=i\} \tag{1}$$

be the transition probability from state i to state j in time t. The corresponding system of forward differential equations is, for $j\leq i$,

$$\begin{aligned} p'_{ij}(t)&=-c\lambda p_{ij}(t)+c\lambda p_{i,j+1}(t), \text{ if } j\geq c\\ p'_{ij}(t)&=-j\lambda p_{ij}(t)+(j+1)\lambda p_{i,j+1}(t), \text{ if } j<c. \end{aligned} \tag{2}$$

A solution to this system is developed in [9]. It is simpler, however, to synthesize the transition probabilities based on the nature of the process. Thus, if $j\leq i\leq c$ we are in the familiar situation of j failures (that is, failure to complete service) in i Bernoulli trials with $e^{-\lambda t}$ as the probability that a customer's service is not completed in time t. It follows that the solution is the well-known binomial

$$p_{ij}(t)=\binom{i}{j}e^{-\lambda tj}(1-e^{-\lambda t})^{i-j} \text{ for } j\leq i\leq c \tag{3}$$

On the other hand, if $i>c$ then there are $i-c$ customers in the center whose service does not begin at time zero. One of these will commence processing as soon as a service channel becomes available prior to time t, etc. If $j\geq c$ then all channels are busy throughout the period and the entire center behaves as a Poisson process with parameter $c\lambda$. Thus

$$p_{ij}(t)=e^{-c\lambda t}\frac{(c\lambda t)^{i-j}}{(i-j)!} \text{ for } c\leq j\leq i. \tag{4}$$

The remaining case, $i>c>j$, is nontrivial but can be handled by familiar methods. Now it is necessary that the system pass through the state c and this can occur at any time u prior to t. In the remaining time $t-u$ the system must go from state c to state j and the required event can be described by taking the union over $0\leq u\leq t$. Thus, by letting p denote the initial number awaiting service,

(5) $$p=i-c,\ p>0,$$

we have

(6) $$p_{ij}(t)=\int_0^t f_p(u)p_{cj}(t-u)\,du \text{ for } i>c\geqslant j,$$

where

$$f_p(u)=\frac{c\lambda}{(p-1)!}(c\lambda u)^{p-1}e^{-c\lambda u},\ u>0$$

is the density of the time required to service p customers (i.e., the first passage time from state i to state c), and $p_{c,j}\ (t-u)$ may be obtained from (3).

The processes described by Eqs. (3) and (4) are well understood and will not be treated further. Before proceeding, it is convenient to introduce N_t, the process that counts the number of departures from the system in time t. Clearly, N_t must satisfy the identity

(8) $$N_t+X_t\equiv i.$$

MOMENTS OF THE COUNTING PROCESS

The probability generating function of N_t can be characterized as follows:

(9) $$\begin{aligned}G(s)=E\{s^{N_t}\}&=\sum_{r=0}^{i}s^r p_{i,\,i-r}(t)=\sum_{r=0}^{i}s^{i-r}p_{i,\,r}(t)\\&=\sum_{r=0}^{c}s^{i-r}\int_0^t f_p(u)p_{cr}(t-u)\,du+\sum_{r=c+1}^{i}s^{i-r}p_{i,\,r}(t).\end{aligned}$$

Using (3) and (4) in the above, summing the binomial, and representing the Poisson sum as an integral results in:

(10) $$\begin{aligned}G(z)=&\,e^{-c\lambda t(1-z)}\int_{tz}^{\infty}f_p(u)\,du\\&+z^p\int_0^t f_p(t-u)\left[e^{-\lambda u}+z(i-e^{-\lambda u})\right]^c du.\end{aligned}$$

The first two moments are of particular interest. They may be found by differentiating (10), setting $z=1$, and reducing. Thus,

(11) $$H(t)=E\{N_t\}=c\lambda t\int_t^{\infty}f_{p-1}(u)\,du+i\int_0^t f_p(u)\,du-c\int_0^t f_p(t-u)e^{-\lambda u}du$$

and

(12) $$\begin{aligned}\Psi(t)=E\{N_t^2\}=&\,(c\lambda t)^2\int_t^{\infty}f_{p-2}(u)\,du+(c\lambda t)\int_t^{\infty}f_{p-1}(u)\,du\\&+i^2\int_0^t f_p(u)\,du-c(2i-1)\int_0^t f_p(t-u)e^{-\lambda u}du\\&+c(c-1)\int_0^t f_p(t-u)e^{-2\lambda u}du.\end{aligned}$$

Expressions (11) and (12) are difficult to work with. For purposes of analysis and of interpreting these moments when p and c are large, it is convenient to work with their Laplace transforms. Thus,

(13) $$H^*(s) = \frac{c\lambda}{s^2}\left\{1 - \left(\frac{c\lambda}{s+c\lambda}\right)^p\right\} + \left\{\frac{c\lambda}{s+c\lambda}\right\}^p \left\{\frac{c}{s} - \frac{c}{s+\lambda}\right\}$$

and

(14) $$\psi^*(s) = \left\{\frac{2(c\lambda)^2}{s^3} + \frac{c\lambda}{s^2}\right\}\left\{1 - \left(\frac{c\lambda}{s+c\lambda}\right)^p\right\} + \left(\frac{c\lambda}{s+c\lambda}\right)^p \left\{\frac{-2pc\lambda}{s^2} + \frac{i^2-p^2}{s} - \frac{c(2i-1)}{s+\lambda} + \frac{c(c-1)}{s+2\lambda}\right\}$$

Suppose c is large, p arbitrary, and θ is their ratio, i.e.,

(15) $$\theta = p/c,$$

then we have approximately

(16) $$\left(\frac{c\lambda}{s+c\lambda}\right)^p \sim e^{-s\theta/\lambda}.$$

Making this substitution in (13) and (14) leads to the approximate moments

(17) $$H^*(s) \sim \frac{c\lambda}{s^2} - e^{-s\theta/\lambda}\left\{\frac{c\lambda}{s^2} - \frac{c}{s} + \frac{c}{s+\lambda}\right\}$$

and

(18) $$\psi^*(s) \sim \frac{2(c\lambda)^2}{s^2} + \frac{c\lambda}{s^2} - e^{-s\theta/\lambda}\left\{\frac{2(c\lambda)^2}{s^3} + \frac{c\lambda}{s^2}(2p+1) - \frac{i^2-p^2}{s} + \frac{c(2i-1)}{s+\lambda} - \frac{c(c-1)}{s+2\lambda}\right\}.$$

The expressions in (17) and (18) can be inverted easily, yielding the following approximations for the mean and variance functions:

(19) $$E\{N_t\} \sim c\lambda t, \text{ for } 0 \leq \lambda t \leq \theta \text{ and } E\{N_t\} \sim p + c[1 - e^{-(\lambda t - \theta)}], \text{ for } \lambda t > \theta$$

and

(20) $$\operatorname{Var}\{N_t\} \sim c\lambda t \text{ for } 0 \leq \lambda t \leq \theta \text{ and } \operatorname{Var}\{N_t\} \sim p + ce^{-(\lambda t - \theta)}[1 - e^{-(\lambda t - \theta)}], \text{ for } \lambda t > \theta.$$

The approximate moments (19) and (20) will be used in developing the central limit theorem for N_t. In applications, it is presumably more desirable to use the exact moments, especially if p is small. These may be obtained either by integrating (11) and (12) or by inverting (13) and (14). The results may be expressed as

(21) $$H(t) = c\lambda t e^{-c\lambda t}\sum_{r=0}^{p-2}\frac{(c\lambda t)^r}{r!} + i\left\{1 - e^{-c\lambda t}\sum_{r=0}^{p-1}\frac{(c\lambda t)^r}{r!}\right\} - c\left(\frac{c}{c-1}\right)^p e^{-\lambda t}\left\{1 - e^{-\lambda t(c-1)}\sum_{r=0}^{p-1}\frac{[\lambda t(c-1)]^r}{r!}\right\}$$

for $c > 1$, $p > 1$ and

(22) $$\begin{aligned}\psi(t) = {} & (c\lambda t)^2 e^{-c\lambda t}\sum_{r=0}^{p-3}\frac{(c\lambda t)^r}{r!} + c\lambda t e^{-c\lambda t}\sum_{r=0}^{p-2}\frac{(c\lambda t)^r}{r!} \\ & + i^2\left\{1 - e^{-c\lambda t}\sum_{r=0}^{p-1}\frac{(c\lambda t)^r}{r!}\right\} \\ & - c(2i-1)\left(\frac{c}{c-1}\right)^p e^{-\lambda t}\left\{1 - e^{-\lambda t(c-1)}\sum_{r=0}^{p-1}\frac{[\lambda t(c-1)]^r}{r!}\right\}\end{aligned}$$

$$+c(c-1)\left(\frac{c}{c-2}\right)^{p} e^{-2\lambda t}\left\{1-e^{-\lambda t(c-2)}\sum_{r=0}^{p-1}\frac{[\lambda t(c-2)]^{r}}{r!}\right\}$$

for $c>2$, $p>2$. (The omitted cases can be obtained from (10) easily.)

CENTRAL LIMIT THEOREMS

Let us introduce the holding time random variable, T_r, representing the time required to service r customers. Clearly,

$$\{X_t>j\}=\{N_t<i-j\}=\{T_{i-j}>t\}. \tag{23}$$

The random variable, T_{i-j}, can be represented as a sum of independent random variables. They will not be identically distributed and we proceed to show that the central limit theorem is valid for them.

Let S_r represent the time between the $r-1$st and the rth departure from the system. Each S_r obeys an exponential probability law; in fact, the probability density is given by

$$f_r(t)=c\lambda e^{-c\lambda t} \text{ for } 1\leq r\leq p, \tag{24}$$

$$f_r(t)=\lambda(i+1-r)e^{-\lambda(i+1-r)t} \text{ for } p+1\leq r\leq i-j \tag{25}$$

The time required to reach state j will be

$$T_{i-j}=\sum_{r=1}^{i-j} S_r \tag{26}$$

and we have

$$m=E\{T_{i-j}\}=\frac{p}{c\lambda}+\frac{1}{\lambda}\sum_{r=j+1}^{c}\frac{1}{r} \tag{27}$$

and

$$\sigma^2=\text{Var }\{T_{i-j}\}=\frac{p}{(c\lambda)^2}+\frac{1}{\lambda^2}\sum_{r=j+1}^{c}\frac{1}{r^2}. \tag{28}$$

A central limit theorem for T_{i-j}, the time to state j, will require that $i-j$ becomes large without limit. But this can occur in several ways. Writing $i-j=p+c-j$ will enable us to examine the choices. If $p\to\infty$ with c and j bounded, then the time spent in transition from state c to state j will be negligible compared to the whole and this situation is equivalent to normal approximation in the identically distributed case (see (27) and (28)). Thus, something new will require (at least) that $c\to\infty$. Next notice that this part of the transition time is negligible also when $j\sim c$ as its contribution to the mean (27) and variance (28) becomes null. Thus the ratio j/c must be bounded below unity. To combine these two thoughts, write $p\sim c\theta$ and $j\sim c\delta$ and consider

$$i-j=p+c-j\sim c(\theta+1-\delta) \tag{29}$$

which will grow without limit unless we have simultaneously $\theta=0$ and $\delta=1$. The case $\theta=0$ has some independent interest, so we will require $\delta<1$.

Now let $f^*(s)$ be the Laplace transform of the probability density of T_{i-j}. It is easily shown that

$$f^*(s)=\left(\frac{c\lambda}{s+c\lambda}\right)^{p}\prod_{r=j+1}^{c}\frac{\lambda r}{s+\lambda r} \tag{30}$$

and hence that

$$\ln f^*(s) = -p \ln \left(1+\frac{s}{c\lambda}\right) - \sum_{r=j+1}^{c} \ln \left(1+\frac{s}{\lambda r}\right). \tag{31}$$

Introducing the normalized random variable

$$U = \frac{T_{i-j}-m}{\sigma} \tag{32}$$

and letting $g^*(s)$ be the Laplace transform of its density function, we note that it is sufficient to show that $\ln g^*(s)$ converges to $s^2/2$.

Using the limited expansion

$$-\ln(1+x) = -x + \frac{x^2}{2} - \frac{1}{3}\frac{x^3}{(1+vx)^3} \text{ and } |v|<1, \tag{33}$$

we have, by use of (27), (28), (32), and (33),

$$\ln g^*(s) = \frac{s^2}{2} + R_2,$$

where

$$R_2 = \frac{-p}{3(c\lambda)^3}\left(\frac{s}{\sigma}\right)^3\left(1+v\frac{s}{c\lambda\sigma}\right)^{-3} - \frac{1}{3}\left(\frac{s}{\sigma}\right)^3 \sum_{r=j+1}^{c} \frac{1}{(\lambda r)^3}\left(1+v\frac{s}{c\lambda\sigma}\right)^{-3}.$$

The factors in (34) that are raised to the power minus three are asymptotically one as $c\to\infty$, so we replace them and express the resulting estimate as

$$|R_2| \leqslant \frac{s^3}{3\sqrt{c}} \frac{\left[\frac{p}{c}+c^2\sum_{j+1}^{c}\frac{1}{r^3}\right]}{\left[\frac{p}{c}+c\sum_{j+1}^{c}\frac{1}{r^2}\right]^{3/2}}, \tag{35}$$

which does not depend on λ.

We require conditions under which the bound (35) converges to zero. By using $p/c=\theta$, it is easy to show that $|R_2|$ converges to zero as $\theta\to\infty$ regardless of the behavior of $j(c\geqslant j)$. This too is equivalent to the identically distributed case. Theorem 1 shows that the normal approximation is valid for arbitrary θ. We note in passing that for $\theta<\infty$ and $\delta<1$ we have $m\to\infty$ and σ^2 converging, (27) and (28).

THEOREM 1: The distribution of $(T_{i-j}-m)/\sigma$ converges to the standard normal distribution as $c\to\infty$ if

$$\begin{aligned} &\text{a) } p/c\to\theta, && 0\leqslant\theta\leqslant\infty \\ &\text{b) } j/c\to\delta, && 0\leqslant\delta<1. \end{aligned} \tag{36}$$

PROOF: It sufficies to show $|R_2|$ converges to zero. Let $\zeta_n(\tau)$ be defined by

$$\sum_{r=1}^{n}\frac{1}{r^\tau} = \frac{-1}{\tau-1}\frac{1}{n^{\tau-1}}+\zeta_n(\tau) \tag{37}$$

for $\tau>1$, and note that $\zeta_n(\tau)$ converges increasingly to $\zeta(\tau)$, the zeta function, as $n\to\infty$. Clearly,

$$\sum_{j+1}^{c}\frac{1}{r^\tau} = \frac{1}{\tau-1}\left(\frac{1}{j^{\tau-1}}-\frac{1}{c^{\tau-1}}\right)+\zeta_c(\tau)-\zeta_j(\tau). \tag{38}$$

It follows that the bound (35) is asymptotically equivalent to

$$\frac{s^3}{3\sqrt{c}} \frac{\left\{\theta\delta^2+\frac{1}{j^2}\,[\zeta(3)-\zeta_j(3)]+\frac{1}{2}\,(1-\delta^2)\right\}}{\left\{\theta\delta+\frac{1}{j}\,[\zeta(2)-\zeta_j(2)]+(1-\delta)\right\}^{3/2}} \tag{39}$$

which, under the conditions $a)$ and $b)$ is

$$0\left(\frac{1}{\sqrt{c}}\right).$$

We note that the case $p\equiv 0$ is covered by the theorem. Thus, as a by-product, we have the central limit theorem for the waiting time to $c-j$ failures in the simple death process.

The normal approximation is also valid for the counting process N_t. This too can occur in many ways. Looking at the approximate moments (19) and (20) we see that if λt (although large) is smaller than θ, then N_t is effectively a Poisson variable and the normal approximation applies. On the other hand, if λt is much larger than θ, which is also large, then the system will be effectively emptied and N_t-c is essentially a Poisson variable. For the remaining cases, the central limit theorem can be established using Theorem 1 and normalizing sequences based on the approximate moments (19) and (20). In addition to hypotheses $a)$ and $b)$, we require a condition concerning the behavior of t, which we now develop. Using $a)$ and $b)$ in (27) and (28), we see that

$$\begin{aligned} m &\sim \frac{1}{\lambda}\{\theta-\log\delta\} \\ \sigma^2 &\sim \frac{1}{j\lambda^2}\{\theta\delta+1-\delta\}. \end{aligned} \tag{38}$$

Using Theorem 1, we can write

$$\Pr\{T_{i-j}>t\}\sim\Pr\left\{\frac{T_{i-j}-m}{\sigma}>\sqrt{j}\,\frac{\lambda t-\theta+\ln\delta}{[\theta\delta+1-\delta]^{1/2}}\right\}\to 1-\Phi(x) \tag{39}$$

where

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-u^2/2}du \tag{40}$$

and this yields the required condition

$$c)\qquad \sqrt{j}\,\frac{\lambda t-\theta+\ln\delta}{[\theta\delta+1-\delta]^{1/2}}\sim x,\ \text{for some } x. \tag{41}$$

Further discussion of this point appears in the next section.

THEOREM 2: Under hypotheses $a)$, $b)$, and $c)$ we have

$$\Pr\{N_t<i-j\}\sim\Phi\left\{K(\delta,\theta)\,\frac{i-j-E(N_t)}{[\mathrm{Var}\,(N_t)]^{1/2}}\right\}, \tag{42}$$

where

$$K(\delta,\theta)=\left\{\frac{\theta+\delta(1-\delta)}{\theta\delta^2+\delta(1-\delta)}\right\}^{1/2}. \tag{43}$$

Either exact (21), (22) or approximate (19), (20) moments may be used in (43). The continuity correction (replace $i-j$ with $i-j-1/2$) may be used because Var $(N_t) \to \infty$.

PROOF: By use of (39) we have

$$\Pr\{N_t < i-j\} = \Pr\{T_{i-j} > t\} \sim 1-\Phi(x). \tag{44}$$

From c) we have

$$Q = e^{-(\lambda t-\theta)} \sim \delta \exp\left\{-x\sqrt{\frac{\theta\delta+1-\delta}{j}}\right\}, \tag{45}$$

where the left hand side of the equivalence relation in (45) serves to define Q. Because $\delta < 1$ there is no loss in taking $\lambda t > \theta$. Thus, using (19) and (20) in (42) one may rewrite the argument as

$$K(\delta, \theta)\frac{i-j-p-c(1-Q)}{\sqrt{p+cQ(1-Q)}} \tag{46}$$

and by virtue of (39) and $\{T_{i-j} > t\} = \{N_t < i-j\}$, the theorem will be proved if expression (46) converges to $-x$. Using conditions b) and (45), we have

$$\begin{aligned} cQ &\sim je^{-x\sqrt{\frac{\theta\delta+1-\delta}{j}}} \\ Q &\sim \delta, \end{aligned} \tag{47}$$

and together with (5) and (47) we can rewrite (46) as

$$K(\delta, \theta)\frac{i-j-p-c(1-Q)}{\sqrt{p+cQ(1-Q)}} \sim -K(\delta, \theta)\frac{j-cQ}{\sqrt{c\theta+c\delta(1-\delta)}} \tag{48}$$

$$\sim -K(\delta, \theta)\sqrt{\theta}\,\frac{\sqrt{j}\left[1-e^{-x\sqrt{\frac{\theta\delta+1-\delta}{j}}}\right]}{\sqrt{\theta+\delta(1-\delta)}} \sim -x, \tag{48}$$

using (43) and (45), as was to be shown.

In reviewing the proofs of Theorems 1 and 2, we see that $\delta = 1$ is allowed if $\theta > 0$. We note from (43) that either $\delta = 1$ and $\theta > 0$, or $\delta < 1$ and $\theta = 0$ will make $K(\delta, \theta) = 1$.

COMPUTATIONAL PROBLEMS

Probabilities for this process have been tabled in [7] for the smaller values of c. Tabular form is not suitable, however, if this process appears as part of a more complex system.

In the project described by Collins and Guthrie [3], the probabilities were obtained by the Runge-Kutta numerical solution of the system of forward differential equations (2). This is reported as satisfactory for $c=6$, $i \leqslant 13$. Lanman [9] found this to be inadequate for $c=8$, $i=16$ and presented a solution to the differential equations resulting in the formula, using $q=c-j$, $p=i-c$, $i > c \geqslant j$,

$$p_{ij}(t) = \sum_{n=j}^{c-1}(-1)^{n+j}\binom{c}{n}\binom{n}{j}\left\{\left(\frac{c}{c-n}\right)^p e^{-n\lambda t} - e^{-c\lambda t}\sum_{r=0}^{p}\frac{(c\lambda t)^r}{r!}\left(\frac{c}{c-n}\right)^{p-r}\right\}. \tag{49}$$

The insertion of (3) and (7) into (6) yields

$$p_{ij}(t) = \frac{(c\lambda)^p}{(p-1)!}e^{-j\lambda t}\binom{c}{j}\int_0^t u^{p-1}\,(e^{-\lambda u}-e^{-\lambda t})^{c-j}du, \tag{50}$$

and (49) can be obtained from (50) upon using a binomial expansion and integrating termwise. Of course such a formula can always be used on a digital computer, but multiple precision arithmetic may be required.

Formula (50) is suitable for analogue computation, but it would be better to reduce the number of parameters. If we make a change of variable so that the upper limit of integration is unity, we see that the product λt may be replaced by λ producing the formula

$$(51)' \qquad p_{ij}(\lambda)=\frac{(c\lambda)^p}{(p-1)!}\binom{c}{j}e^{-j\lambda}\int_0^1 u^{p-1}\,(e^{-\lambda u}-e^{-\lambda})^q du$$

with a minor change in notation. We note in passing that some experimentation was performed to see if the method of Laplace (see e.g., [4]) could be used to approximate the integral in (50). The method is not adequate.

The following formula (54) has proved useful for small λ. Rewrite (50) as

$$(52) \qquad P_{ij}(\lambda)=\frac{(c\lambda)^p}{(p-1)!}\binom{c}{j}e^{-c\lambda}\int_0^1 u^{p-1}(e^{\lambda(1-u)}-1)^q du$$

and expand the integral portion as follows

$$(53) \qquad (-1)^q\sum_{n=0}^{q}\binom{q}{n}(-1)^n\int_0^1 u^{p-1}e^{n\lambda(1-u)}=(-1)^q\sum_{n=0}^{q}\binom{q}{n}(-1)^n\sum_{k=0}^{\infty}\frac{(n\lambda)^k}{k!}\frac{\Gamma(p)\Gamma(k+1)}{\Gamma(p+k+1)},$$

and obtain the ultimate formula

$$(54) \qquad p_{ij}(\lambda)=(c\lambda)^p\binom{c}{q}e^{-c\lambda}(-1)^q\sum_{n=1}^{q}(-1)^n\binom{q}{n}\sum_{k=1}^{\infty}\frac{(n\lambda)^k}{(p+k)!}.$$

The infinite series in (54) is dominated by

$$\sum_{k=1}^{\infty}\frac{(q\lambda)^k}{(p+k)!},$$

and the error after $M-1$ terms may be estimated by

$$(55) \qquad \sum_{M}^{\infty}\frac{(q\lambda)^k}{(p+k)!}\leqslant\frac{(q\lambda)^M}{(p+M)!}\sum_{0}^{\infty}\frac{(q\lambda)^k}{k!}=\frac{(q\lambda)^M}{(p+M)!}e^{q\lambda}.$$

For large values of c, approximations based on the normal distribution can be used. Two questions arise. Which approximation is best, and how large must c be? If the process is essentially Poisson (i.e., the service center is always full with high probability), or essentially binomial (i.e., the service center is not full with high probability) then the storehouse of practical experience for the normal approximation to those two distributions may be used. Otherwise, we can use the approximations given herein. To this end we consider the questions of the values of c and θ for which this may be appropriate.

Since the Poisson and binominal aspects are intermingled, consider the moments in (19) and (20) for $\lambda>\theta$. (We have already shown that t may be set equal to one without loss.) Let us require that the contribution of the binomial to the variance represents at least a multiple α of that of the Poisson. That is, using Q from (45) and $P=1-Q$, we look for parameterizations for which

$$(56) \qquad \operatorname{Var}(N_t)\sim p+cPQ\geqslant p(1+\alpha).$$

We note that the same requirement is then automatically placed on the mean because $cP \geqslant cPQ$. The inequality (56) defines the region in the θ, λ plane given by

$$\ln \frac{2}{1+\sqrt{1-4\alpha\theta}} \leqslant \lambda-\theta \leqslant \ln \frac{2}{1-\sqrt{1-4\alpha\theta}}. \tag{57}$$

A plot of this region appears in Figure 1, where $\alpha\theta$ appears on the horizontal and $\lambda-\theta$ on the vertical. It may be used to help choose among the various normal approximations. For $i=20$, $c=10$, $\lambda=1.0$, we found that Theorems 1 and 2 both gave values valid within a magnitude of 0.05. The use of a continuity correction in Theorem 2 seems to improve the approximation. Practitioners in renewal theory seem to prefer the Theorem 1 approach for computation.

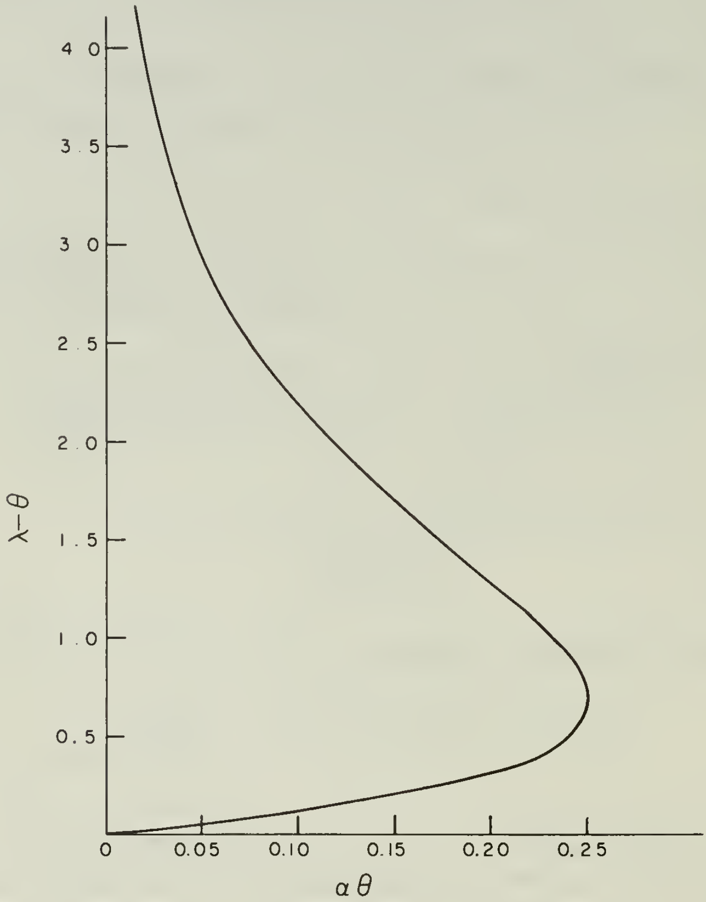


Figure 1. The region of interesting parameter values.

APPLICATION TO PROBLEMS IN LIFE TESTING

The life testing procedures introduced by Epstein and Sobel [5], [6] use the exponential distribution to advantage in test plans designed to terminate after either a fixed number of failures or fixed time (as opposed to a fixed number of items tested). They show that the total test time (of all subsystems) is sufficient for λ. They were not concerned with idle time of test equipment. Also, the case of a limited number i of subsystems available for test and a lesser number c of test chambers has not been treated. The distribution theory follows easily from the results already presented.

For purposes of laying a base, it is convenient to treat the case $i=c$ (i.e., $p=0$) first and expand from this later. We start with the stopping rule based on a fixed number r of failures. Let T_r, B_r, I_r be (respectively) the time to r failures, the total test time of the first r to fail, and the total service center idle time during the first r failures. Clearly,

$$T_r=\sum_{j=1}^{r} S_j,\ B_r=\sum_{j=1}^{r} (c-j+1)S_j,\ I_r=\sum_{j=1}^{r} (j-1)S_j \tag{58}$$

and because of the constraint

$$cT_r=B_r+I_r, \tag{59}$$

the joint distribution of T_r, B_r, I_r is contained in a plane. Let the transform of this joint density be

$$f^*(u, v, w)=E\{e^{-uT_r-vB_r-wI_r}\}. \tag{60}$$

This may be deduced from the joint transform of $S_1, S_2, \ldots, S_r$ as follows. From (25) and (30), we see that this latter transform is

$$E\left\{e^{-\sum_1^r s_jS_j}\right\}=\prod_{j=1}^{r}\frac{\lambda(c-j+1)}{s_j+\lambda(c-j+1)}. \tag{61}$$

Then, by using (58) it is seen that

$$\sum_1^r s_jS_j=uT_r+vB_r+wI_r \tag{62}$$

if

$$s_j=u+(c-j+1)v+(j-1)w,\ \text{for } j=1, \ldots, r. \tag{63}$$

Inserting (63) into (61) yields

$$f^*(u, v, w)=\lambda^r\frac{c!}{(c-r)!}\prod_{j=1}^{r}[u+(v+\lambda)(c-j+1)+(j-1)w]^{-1}. \tag{64}$$

Expression (64) can be inverted to obtain any of the bivariate distributions T_r, B_r or T_r, I_r or B_r, I_r. To illustrate, we'll treat the case T_r, B_r which perhaps has the greatest interest in questions of experimental design.

Setting $w=0$ in (64) and expanding that expression in partial fractions yields

$$\begin{aligned}E\{e^{-uT_r-vB_r}\}=f^*(u,v,0)&=\iint e^{-ut-vb}f(t,b)\,dt\,db\\ &=\frac{r\lambda^r}{(\lambda+v)^{r-1}}\binom{c}{r}\sum_{j=1}^{r}(-1)^{r-j}\binom{r-1}{j-1}\frac{1}{[u+(\lambda+v)(c-j+1)]}.\end{aligned} \tag{65}$$

Holding v fixed and inverting with respect to u results in

$$\int_0^\infty e^{-vb}f(t,b)\,db=\frac{r\lambda^r}{(\lambda+v)^{r-1}}\binom{c}{r}\sum_{j=1}^{r}(-1)^{r-j}\binom{r-1}{j-1}e^{-t(v+\lambda)(c-j+1)}. \tag{66}$$

We take this opportunity to obtain the marginal distribution of T_r. The series can be summed by the binomial theorem and upon setting $v=0$, we have

$$f(t)=r\lambda\binom{c}{r}e^{-\lambda t(c-r+1)}[1-e^{-\lambda t}]^{r-1},\ t>0. \tag{67}$$

The marginal distribution of B_r is the r-fold convolution of the exponential distribution with parameter λ, e.g., see $f^*(0, v, 0)$, (64).

Returning to (66), we can invert termwise. Each term has the form

$$\frac{1}{(\lambda+v)^{r-1}} e^{-(\lambda+v)\alpha}, \tag{68}$$

whose inverse is

$$\frac{(b-\alpha)^{r-2}}{(r-2)!} e^{-b\lambda} \cup (b-\alpha), \tag{69}$$

where $\cup(x)$ is the unit step function:

$$\begin{aligned} \cup(x) &= 1 \quad \text{for } x \geq 0 \\ &= 0 \quad \text{for } x < 0 \end{aligned} \tag{70}$$

It follows that the joint density of T_r and B_r is

$$f(t, b) = r\lambda^r \binom{c}{r} \sum_{j=0}^{r-1} (-1)^{r-j-1} \binom{r-1}{j} e^{-b\lambda} \frac{[b-t(c-j)]^{r-2}}{(r-2)!} \cup (b-t(c-j)). \tag{71}$$

Of course the remaining bivariate distributions can be deduced using (59).

For large r the multivariate normal approximation is valid. The pattern of Theorem 1 will be followed in showing this. We need the first and second order moments which can be obtained by differentiating the log of the transform (64). Thus

$$ln\, f^*(u, v, w) = r\, ln\, \lambda + ln \frac{c!}{(c-r)!} - \sum_{j=0}^{r-1} ln\, [u + (v+\lambda)(c-j) + jw], \tag{72}$$

from which we calculate the mean vector

$$\begin{aligned} E(T_r) &= \frac{1}{\lambda} \sum_{j=0}^{r-1} \frac{1}{c-j} \\ E(B_r) &= \frac{r}{\lambda} \\ E(I_r) &= \frac{1}{\lambda} \sum_{j=1}^{r-1} \frac{j}{c-j} \end{aligned} \tag{73}$$

and the covariance matrix

$$\begin{aligned} \sigma^2(T_r) &= \frac{1}{\lambda^2} \sum_{j=0}^{r-1} \frac{1}{(c-j)^2} & \sigma(T_r, B_r) &= \frac{1}{\lambda^2} \sum_{0}^{r-1} \frac{1}{c-j} \\ \sigma^2(B_r) &= \frac{r}{\lambda^2} & \sigma(B_r, I_r) &= \frac{1}{\lambda^2} \sum_{1}^{r-1} \frac{j}{(c-j)} \\ \sigma^2(I_r) &= \frac{1}{\lambda^2} \sum_{j=1}^{r-1} \left(\frac{j}{c-j}\right)^2 & \sigma(T_r, I_r) &= \frac{1}{\lambda^2} \sum_{1}^{r-1} \frac{j}{(c-j)^2}. \end{aligned} \tag{74}$$

Since r indexes the number of failures, the state of the system is $c-r$. It is this latter quantity to which condition b) refers, so we rewrite it as condition d).

THEOREM 3: Under the condition

$$d)\ \frac{r}{c} \to 1-\delta \text{ with } 0 \leq \delta < 1 \text{ as } c \to \infty$$

the asymptotic distribution of the vector (T_r, B_r, I_r) is normal with mean vector given by (73) and covariance matrix given by (74).

PROOF: Following the proof of Theorem 1, we apply the limited expansion (33) to

$$\log f^*(u, v, w) = -\sum_{j=0}^{r-1} \log \left\{1 + \frac{v}{\lambda} + \frac{u+jw}{\lambda(c-j)}\right\}. \tag{75}$$

The linear terms will disappear upon removal of the mean. The quadratic terms have (74) for their matrix and this will transform to the quadratic form in correlations when we replace u with $u/\sigma(T_r)$, v with $v/\sigma(B_r)$, and w with $w/\sigma(I_r)$. Applying (37) and condition $d)$ to the variances in (74) permits us to write

$$\begin{aligned} \sigma^2(T_r) &\sim c(1-\delta)/\delta\lambda^2 \\ \sigma^2(B_r) &\sim c\delta/\lambda^2 \\ \sigma^2(I_r) &\sim c\{2\delta \log \delta + (1-\delta)(1+\delta)\}/\delta\lambda^2, \end{aligned} \tag{76}$$

which are used to estimate the error term in the expansion for the normalized transform. Thus,

$$\begin{aligned} |R_2| &\leq \frac{1}{3\lambda^3} \sum_{j=0}^{r-1} \left| \frac{v}{\sigma(B_r)} + \frac{1}{c-j}\left[\frac{u}{\sigma(T_r)} + j\frac{w}{\sigma(I_r)}\right]\right|^3 \\ &\leq \frac{1}{3}\sqrt{\delta/c}\left| v + \frac{u}{\sqrt{1-\delta}} + \frac{w(1-\delta)}{\sqrt{2\delta \ln \delta + (1-\delta)(1+\delta)}}\right|^3 \end{aligned} \tag{77}$$

which is $0(1/\sqrt{c})$ and the theorem is proved.

Let us turn now to stopping rules based on fixed time t, making the number of failures random. Let N_t, B_t, I_t represent (respectively) the number of failures, the total test time, and the total idle time in the period $(0, t)$. Since $(t - T_{N_t})$ is the time since the most recent failure, we have

$$\begin{aligned} B_t &= B_{N_t} + (t - T_{N_t})(c - N_t) \\ I_t &= I_{N_t} + (t - T_{N_t})N_t \end{aligned} \tag{78}$$

and upon using (59) in (78) it follows that

$$B_t + I_t = ct, \tag{79}$$

a result which, in itself, is obvious. From (79) we see that it is sufficient to work with either B_t or I_t, and (78) suggests that the latter choice is simpler. We proceed to develop the joint distribution of N_t and I_t.

From (78) and (23) we see that

$$\Pr\{N_t = r, I_t \leq z\} = \Pr\{T_r \leq t, T_{r+1} > t, I_r + (t-T_r)r \leq z\} = \int_0^t dx \int_t^\infty dy \int_0^z dz' f(x, y, z'), \tag{80}$$

where $f(x, y, z)$ is the joint density of T_r, T_{r+1}, J_r and

$$J_r = I_r - rT_r + rt. \tag{81}$$

Let the transform of the density be

$$f^*(u, v, w) = e^{-wrt} E\{e^{-uT_r - vT_{r+1} - w(J_r - rt)}\} \tag{82}$$

and note that, by using (81) and (58),

(83)
$$\sum_1^{r+1} s_j S_j = uT_r + vT_{r+1} + w(J_r - rt)$$

if

(84)
$$s_j = u + v - (r-j+1)w \qquad \text{for } j = 1, \ldots, r$$
$$s_{r+1} = v.$$

Making this substitution in (61) yields

(85)
$$f^*(u, v, w) = e^{-wrt} \frac{\lambda(c-r)}{v+\lambda(c-r)} \prod_{j=1}^{r} \frac{\lambda(c-j+1)}{[u+v-(r-j+1)w+(c-j+1)\lambda]}$$

and this contains the case $r=0$ because then $I_t \equiv 0$ and the product over an empty set is unity. The plan of attack is to invert (85) first with respect to u and then v. Then it will be convenient to integrate with respect to x and y as described in (80). This will be followed by inversion with respect to w.

Now, the product portion of (85) has the partial fraction expansion

(86)
$$r\lambda^r \binom{c}{r} \frac{1}{(w-\lambda)^{r-1}} \sum_{j=0}^{r-1} (-1)^j \binom{r-1}{j} \frac{1}{u+v+(c-j)\lambda-(r-j)w}$$

and inversion of (86) with respect to u produces

(87)
$$\frac{r\lambda^r}{(w-\lambda)^{r-1}} \binom{c}{r} e^{-x(v+c\lambda-rw)} \sum_{j=0}^{r-1} (-1)^j \binom{r-1}{j} e^{-jx(w-\lambda)}$$
$$= \frac{r\lambda^r}{(w-\lambda)^{r-1}} \binom{c}{r} e^{-x[v+\lambda(c-r)-r(w-\lambda)]} \{1-e^{-x(w-\lambda)}\}^{r-1}.$$

Next, gather the factors involving v

(88)
$$\frac{1}{v+(c-r)\lambda} e^{-x[v+(c-r)\lambda]}$$

and invert, thus producing

(89)
$$e^{-\lambda y(c-r)} \cup (y-x).$$

After integrating with respect to x and y, let's collect the results obtained so far:

(90)
$$\int_0^t dx \int_t^\infty dy \int_0^\infty dz e^{-wz} f(x, y, z)\, dx dy dz$$
$$= e^{-wrt} e^{-\lambda t(c-r)} \frac{\lambda^r}{(w-\lambda)^r} \binom{c}{r} [e^{t(w-\lambda)} - 1]^r$$
$$= e^{-c\lambda t} \binom{c}{r} \lambda^r \{[1-e^{-t(w-\lambda)}]/(w-\lambda)\}^r.$$

Finally, we invert on w to obtain

(91)
$$\frac{d}{dz} \Pr\{N_t = r, I_t \leq z\} = \int_0^t dx \int_t^\infty (x, y, z)$$
$$= e^{-\lambda(ct-z)} \binom{c}{r} \lambda^r \sum_{j=0}^{r} (-1)^j \binom{r}{j} \frac{(z-jt)^{r-1}}{(r-1)!} \cup (z-jt) \qquad \text{for } 0 \leq z \leq ct.$$

The conditional distribution of I_t given $N_t = r$ has an interesting interpretation. Since

$$\Pr\{N_t = r\} = \binom{c}{r} e^{-\lambda t(c-r)}[1-e^{-\lambda t}]^r, \tag{92}$$

we can divide into the transform (90) to produce

$$E\{e^{-wI_t}|N_t = r\} = \left[\frac{\lambda e^{-\lambda t}}{1-e^{-\lambda t}}\right]^r \left[\frac{e^{t(\lambda-w)}-1}{\lambda w}\right]^r. \tag{93}$$

Equation (93) tells us that, given $N_t = r$, the idle time has the distribution as the sum of r truncated exponential variables with density

$$f(z) = \frac{\lambda}{e^{\lambda t}-1} e^{\lambda z},\ 0 \leq z \leq t. \tag{94}$$

Also, by using (92) and summing, we can write

$$E\left\{e^{-uN_t - wI_t}\right\} = e^{-c\lambda t}\left\{1 + \lambda e^{-u}\frac{e^{t(\lambda-w)}-1}{\lambda - w}\right\}^c \tag{95}$$

and the bivariate normal approximation to the distribution of $\{N_t, I_t\}$ can be justified in a straightforward way. Setting $u = 0$ in (95) and inverting produces the marginal density of I_t, namely

$$f(z) = \sum_{r=0}^{c}\binom{c}{r}\frac{\lambda^r}{(r-1)!}e^{\lambda z}\sum_{j=0}^{r}(-1)^j\binom{r}{j}(z-jt)^{r-1}\cup(z-jt) \tag{96}$$
$$\text{for } 0 \leq z \leq \text{ct}.$$

Let's take this opportunity to record the first and second moments, which can be obtained from (95)

$$E\{N_t\} = c\{1-e^{-\lambda t}\} \tag{97}$$
$$E\{I_t\} = \frac{c}{\lambda}\{\lambda t - 1 + e^{-\lambda t}\}$$

$$\mathrm{Var}\{N_t\} = ce^{-\lambda t}\{1-e^{-\lambda t}\} \tag{98}$$
$$\mathrm{Var}\left\{I_t\right\} = \frac{c}{\lambda^2}\left\{1 - 2\lambda t e^{-\lambda t} - e^{-2\lambda t}\right\}$$
$$\mathrm{Cov}\{N_t, I_t\} = \frac{c}{\lambda}e^{-\lambda t}\{\lambda t - 1 + e^{-\lambda t}\}\cdot$$

Now let's turn to the question of extending these results to the case of limited test facilities, i.e., $c < i$ and hence $i - c = p > 0$. Consider first stopping rules based on a fixed number of failures and let T_r, B_r, I_r be defined as before. In this case the triplet will be one dimensional as long as $r \leq p$ because then $B_r = cT_r$ and $I_r \equiv 0$. This process is well understood so we consider only the case $r > p$, and then the triplet becomes two-dimensional again. It is convenient to use the following notation

$$T_r = T_p + T_{p,r}$$
$$B_r = cT_p + B_{p,r} \tag{99}$$
$$I_r = I_{p,r},$$

where $T_{p,r}$, $B_{p,r}$, and $I_{p,r}$ denote (respectively) the time, total test time, and the idle time from the pth failure to the rth failure. Moreover, the process $T_{p,r}$, $B_{p,r}$, $I_{p,r}$ has been studied already. In

particular, the transform (60) can be rewritten

$$f^*(u,v,w)=E\{\exp[-u(T_r)-v(B_r)-w(I_r)]\}$$
$$=E\{\exp[-(u+cv)T_p]\}E\{\exp(-uT_{p,r}-vB_{p,r}-wI_{p,r}\} \tag{100}$$
$$=\left[\frac{\lambda c}{u+cv+\lambda c}\right]^p\prod_{j=1}^{r-p}\frac{\lambda(c-j+1)}{u+(v+\lambda)(c-j+1)+(j-1)w}$$

upon re-interpreting (30) and (64). The mean vector (73) must be replaced by

$$E(T_r)=\frac{p}{c\lambda}+\frac{1}{\lambda}\sum_{j=0}^{r-p-1}\frac{1}{c-j}$$
$$E(B_r)=\frac{r}{\lambda} \tag{101}$$
$$I(I_r)=\frac{1}{\lambda}\sum_{j=1}^{r-p-1}\frac{j}{c-j},$$

and the covariance matrix (74) becomes

$$\sigma^2(T_r)=\frac{p}{(c\lambda)^2}+\frac{1}{\lambda^2}\sum_{j=0}^{r-p-1}\frac{1}{(c-j)^2}\qquad \sigma(T_r,B_r)=\frac{p}{c\lambda^2}+\frac{1}{\lambda^2}\sum_{j=0}^{r-p-1}\frac{1}{c-j}$$
$$\sigma^2(B_r)=\frac{r}{\lambda^2}\qquad \sigma(B_r,I_r)=\frac{1}{\lambda^2}\sum_{j=1}^{r-p-1}\frac{j}{c-j} \tag{102}$$
$$\sigma^2(I_r)=\frac{1}{\lambda^2}\sum_{j=1}^{r-p-1}\left(\frac{j}{c-j}\right)^2\qquad \sigma(T_r,I_r)=\frac{1}{\lambda^2}\sum_{j=1}^{r-p-1}\frac{j}{(c-j)^2}.$$

Theorem 3 can be extended to this case following the pattern of Theorem 1. The details will be omitted, and we state without proof:

THEOREM 4: Under the conditions

(a) $\frac{p}{c}\to\theta$ with $0\le\theta<\infty$

(e) $\frac{r-p}{c}\to 1-\delta$ with $0\le\delta<1$

as $c\to\infty$, the vector (T_r, B_r, I_r) is asymptotically normal with mean given by (101) and covariance matrix (102).

Let us now turn to the case of fixed time with $p>0$. Equation (78) requires a slight modification, namely

$$B_t=ct,\ I_t\equiv 0\qquad \text{if } N_t\le p$$
$$B_t=B_{N_t}+(t-T_{N_t})(c+p-N_t)\qquad \text{if } N_t>p \tag{103}$$
$$I_t=I_{N_t}+(t-T_{N_t})(N_t-p)\qquad \text{if } N_t>p$$

and (79) remains valid. Again it is sufficient to work with I_t.

In order to characterize the joint distribution of N_t and I_t it will be convenient to introduce some further notations. For $t>\tau$ let

$$N_{\tau,t}=N_t-N_\tau \tag{104}$$
$$I_{\tau,t}=I_t-I_\tau.$$

Let us proceed to calculate

$$(105)\qquad E\{e^{-uN_t-wI_t}\}=E\{e^{-uN_t-wI_t}|T_p>t\}\ \Pr\{T_p>t\} + \int_0^t E\{e^{-uN_t-wI_t}|T_p=\tau\}f_p(\tau)\,d\tau,$$

with $f_p(\tau)$ given by (7). From (104) we have for $\tau\leq t$

$$(106)\qquad E\{e^{-uN_t-wI_t}|T_p=\tau\}=E\{e^{-uN_\tau-wI_\tau}|T_p=\tau\}E\{e^{-uN_{\tau,t}-wI_{\tau,t}}|T_p=\tau\} - e^{-up}e^{-c\lambda(t-\tau)}\left\{1+\lambda e^{-u}\frac{e^{(t-\tau)(\lambda-w)}-1}{\lambda-w}\right\}^c$$

by use of the Markov property, (103) and (95). It follows that we have the characterization

$$(107)\qquad E\{e^{-uN_t-wI_t}\}=\sum_{r=0}^{p-1}e^{-ur}e^{-c\lambda t}\frac{(c\lambda t)^r}{r!} + e^{-up}\int_0^t f_p(\tau)\left\{e^{-\lambda(t-\tau)}+\frac{\lambda}{\lambda-w}e^{-u}[e^{-w(t-\tau)}-e^{-\lambda(t-\tau)}]\right\}^c d\tau,$$

where $f_p(\tau)$ is given by (7). From this we can develop the moments

$$(108)\qquad E(I_t)=\frac{c}{\lambda}\int_0^t\{\lambda(t-\tau)-1+e^{-\lambda(t-\tau)}\}f_p(\tau)\,d\tau$$

$$E(I_t^2)=\frac{c}{\lambda^2}\int_0^t\{[\lambda^2(t-\tau)^2-2\lambda(t-\tau)+2(1-e^{-\lambda(t-\tau)})] + (c-1)[\lambda(t-\tau)-(1-e^{-\lambda(t-\tau)})]^2\}f_p(\tau)\,d\tau$$

$$\operatorname{Cov}(N_t,I_t)=\frac{c}{\lambda}\int_0^t\{\lambda(t-\tau)-(1-c^{-\lambda(t-\tau)})\}\{(c-1)[1 - e^{-\lambda(t-\tau)}]-1\}f_p(\tau)\,d\tau-E(I_t)\{E(W_t)-p\}$$

and the first two moments of N_t have already been recorded (11) and (12), or (21) and (22). Approximate moments for large c can be developed in the same way as was done to produce (19) and (20). Namely,

$$(109)\qquad E(I_t)\sim\frac{c}{\lambda}\{\lambda t-\theta-(1-e^{-(\lambda t-\theta)})\}\quad\text{for }\lambda t>\theta$$
$$\sim 0\qquad\text{otherwise}$$

$$\operatorname{Var}(I_t)\sim\frac{c}{\lambda^2}\{1-2(\lambda t-\theta)e^{-(\lambda t-\theta)}-e^{-2(\lambda t-\theta)}\}\ \text{for }\lambda t>0$$
$$\sim 0\qquad\text{otherwise}$$

$$\operatorname{Cov}(N_t,I_t)\sim\frac{c}{\lambda}\{(\lambda t-\theta)-(1-e^{-(\lambda t-\theta)})\}e^{-(\lambda t-\theta)}\ \text{for }\lambda t>\theta$$
$$\sim 0\qquad\text{otherwise.}$$

Although a bivariate central limit theorem for (N_t, I_t) could probably be developed after the pattern of Theorem 2, it has little more than academic interest. Such is not easily proved. The chief difficulty lies in the role played by the function $K(\delta, \theta)$ of (43). For computational purposes, sharper results

can be obtained by using (103) in the expressions

$$\Pr\{N_t \leqslant k,\ I_t \leqslant z\} = \Pr\{N_t \leqslant k\} \quad \text{for } 0 \leqslant k < p \text{ and } z > 0 \tag{110}$$

$$\Pr\{N_t \leqslant p+k,\ I_t \leqslant z\} = \sum_{r=0}^{k} \Pr\{N_t = p+r\}\ \Pr\{I_{p+r} - rT_{p+r} + rt \leqslant z | N_t = p+r\}$$

Theorems 2 and 4 can be used in the right hand side of (110) to produce approximations.

REFERENCES

[1] Barlow, R. E. and F. Proschan, *Mathematical Theory of Probability* (The SIAM Series in Applied Mathematics) (John Wiley and Sons, Inc., New York, 1965).

[2] Burke, P. J., "The Output of a Queueing System," Operations Research, **4**, 699–704 (1956).

[3] Collins, F. C. and D. Guthrie, "A Model for the Analysis of AEW and CAP Aircraft Availability," Nav. Res. Log. Quart., **10**, 73–79 (1963).

[4] deBruijn, N. G., "Asymptotic Methods in Analysis," Interscience (1958).

[5] Epstein, B., "Truncated Life Tests in the Exponential Case," Ann. Math. Stat., **25**, 555–564 (1954).

[6] Epstein, B. and M. Sobel, "Life Testing I," J. Am. Stat. Assoc., **48**, 486–502 (1953).

[7] Heyne, J. B. and L. Brotman, "On Pulse-Type Arrivals to a Queue," Operations Research, **8**, 407–418 (1960).

[8] Kendall, D. G., "Stochastic Processes Occurring in the Theory of Queues and Their Analysis by the Method of the Imbedded Markov Chain," Ann. Math. Stat., **24**, 338–354 (1953).

[9] Lanman, G. M., "An Adaptation of a Markov Chain Model for Antisubmarine Warfare Carrier Aircraft," Masters Thesis, U.S. Naval Postgraduate School, Monterey, Calif. (May 1966), AD489085.

[10] Reich, Edgar, "Waiting Times When Queues Are in Tandem," Ann. Math. Stat., **28**, 768–773 (1957).

[11] Saaty, T. L., Elements of Queueing Theory (McGraw-Hill Book Co., New York, 1961).

[12] Saaty, T. L., "Time Dependent Solution to the Many Server Poisson Queue," Operations Research, **8**, 755–772 (1960).

* * *

NEWS AND MEMORANDA

List of Referees

The Editors of the Naval Research Logistics Quarterly are grateful to the following individuals for assisting in the review of articles prior to publication.

Arthur Albert
C. Ancker
Francis Anscombe
Max Astrachan
R. J. Aumann
Egon Balas
Alfredo Banos
E. W. Barankin
C. B. Barfoot
R. E. Barlow
William J. Baumol
Robert Bechhofer
Richard Bellman
L. D. Berkovits
Stuart Bessler
Z. W. Birnbaum
Wallace R. Blischke
Saul Blumenthal
D. Boyer
Jerome Bracken
Stephen P. Bradley
Joseph Bram
R. G. Brandenburg
A. Charnes
H. Chernoff
W. W. Cooper
R. Cottle
J. M. Cozzolino
L. Cutler
Walter Deemer
John L. Denny, Jr.
Cyrus Derman
D. A. D'Esopo
James Dobbie
Melvin Dresher
Satya D. Dubey
A. J. Duncan
Paul S. Dwyer
Richard Eisenman
Henry Ellner
Salah Elmaghraby
Richard Evans
Tibor Fabian
James Falk
A. M. Feiler
Anthony Fiacco
Thomas Ferguson
S. I. Firstman
Peter Fishburn
Bennett Fox
David Gale
Saul I. Gass
Donald P. Gaver
Murray Geisler
A. J. Goldman
Ralph Gomory
Donald Gross
Shanti S. Gupta
Sheldon Haber
Carl Harris
Leon Harter
Klaus Heiss
Robert Helmbold
Leon Herbach
F. S. Hillier
Warren Hirsch
V. Hodgson
Alan J. Hoffman
John Honig
David Howes
Te Chiang Hu
Donald L. Iglehart
Edward Ignall
Rufus Isaacs
J. R. Jackson
William Jewell
Vernon Johns
Ellis L. Johnson
H. C. Joksch
John H. K. Kao
G. D. Kaye
Morton Klein
Leonard Kleinrock
P. J. Kolesar
Kenneth Kortanek
H. W. Kuhn
Jack Laderman
Joel Levy
Ray Lewis
Gerald J. Lieberman
Steven Liebholz
Robert Love
Robert J. Lundegard
Garth McCormick
Bruce J. McDonald
W. G. Madow
Harry Markowitz
George Marsaglia
Clifford Marshall
Michael Maschler
William L. Maxwell
Frederic Miercort
Joseph Moder
Richard Morey
Oskar Morgenstern
John F. Muth
George Nemhauser
Marcel F. Neuts
Arthur J. Nevins
George Nicholson
Henry Nuttle
Ingram Olkin
William R. Pabst, Jr.
Emanuel Parzen
Ralph W. Pfouts
Narahari U. Prabhu
Frank Proschan
Roy Radner
Fred Rigby
T. Rockafellar
Sheldon M. Ross
Thomas Saaty
Richard Savage
Herbert Scarf
Ray Schafer
F. M. Scherer
Roger Schroeder
David Shanno
L. Shapley
M. L. Shooman
Walter L. Smith
Henry Solomon
Herbert Solomon
Ivar Stakgold
Andrew C. Stedry
George Suzuki
Howard Taylor
Gerald Thompson
William Thompson, Jr.
Che S. Tsao
William F. Tyndall
Richard Van Slyke
Thomas Varley
Arthur F. Veinott, Jr.
H. M. Wagner
K. T. Wallenius
H. Weingarten
Lionel Weiss
Andrew Whinston
W. L. Wilkinson
Robert B. Wilson
Peter Winters
Philip Wolfe
Earl Yost
Shelley Zacks
W. I. Zangwill
Peter Zehna

INDEX OF VOLUME 16

LENEMAN, O. A. Z. and F. J. BEUTLER, "On a New Approach to the Analysis of Stationary Inventory Problems," Vol. 16, No. 1, Mar. 1969, pp. 1–15.

LIEBERMAN, G. J., "The Status and Impact of Reliability Methodology," Vol. 16, No. 1, Mar. 1969, pp. 17–35.

LIENTZ, B. P. and R. ALTER, "Applications of a Generalized Combinatorial Problem of Smirnov," Vol. 16, No. 4, Dec. 1969, pp. 543–547.

LIPPMAN, S. A. and J. S. C. YUAN, "Discounted Production Scheduling and Employment Smoothing," Vol. 16, No. 1, Mar. 1969, pp. 93–110.

LOVE, R. F., "Locating Facilities in Three-Dimensional Space by Convex Programming," Vol. 16, No. 4, Dec. 1969, pp. 503–516.

LU, J. Y. and R. B. S. BROOKS, "War Reserve Spares Kits Supplemented by Normal Operating Assets," Vol. 16, No. 2, June 1969, pp. 229–236.

LUTFUL KABIR, A. B. M. and L. K. CHAN, "Optimum Quantiles for the Linear Estimation of the Parameters of the Extreme Value Distribution in Complete and Censored Samples," Vol. 16, No. 3, Sept. 1969, pp. 381–404.

MANN, NANCY R., "A Test for the Hypothesis That Two Extreme-Value Scale Parameters Are Equal," Vol. 16, No. 2, June 1969, pp. 207–216.

MARSHALL, C. W., "Quantification of Contractor Risk," Vol. 16, No. 4, Dec. 1969, pp. 531–541.

MARSHALL, K. T. and J. W. SUURBALLE, "A Note on Cycling in the Simplex Method," Vol. 16, No. 1, Mar. 1969, pp. 121–137.

MAXWELL, W. L. and T. B. CRABILL, "Single Machine Sequencing with Random Processing Times and Random Due-Dates," Vol. 16, No. 4, Dec. 1969, pp. 549–554.

MAZUMDAR, M., "Uniformly Minimum Variance Unbiased Estimates of Operational Readiness and Reliability in a Two-State System," Vol. 16, No. 2, June 1969, pp. 199–206.

MIDLER, J. L. and R. D. WOLLMER, "Stochastic Programming Models for Scheduling Airlift Operations," Vol. 16, No. 3, Sept. 1969, pp. 315–330.

MORGENSTERN, O. and G. L. THOMPSON, "An Open Expanding Economy Model," Vol. 16, No. 4, Dec. 1969, pp. 443–457.

NEMHAUSER, G. L., M. BELLMORE and W. D. EKLOF, "A Decomposable Transshipment Algorithm for a Multiperiod Transportation Problem," Vol. 16, No. 4, Dec. 1969, pp. 517–524.

RANDOLPH, P. H. and G. E. SWINSON, "The Discrete Max-Min Problem," Vol. 16, No. 3, Sept. 1969, pp. 309–314.

RAY, T. L. and P. L. DAVIS, "A Branch-Bound Algorithm for the Capacitated Facilities Location Problem," Vol. 16, No. 3, Sept. 1969, pp. 331–344.

READ, R. R., "On the Output of Parallel Exponential Service Channels," Vol. 16, No. 4, Dec. 1969, pp. 555–572.

SCHWARTZ, A. N. and A. J. BONESS, "Aircraft Replacement Policies in the Naval Advanced Jet Pilot Training Program: A Practical Example of Decision-Making Under Incomplete Information," Vol. 16, No. 2, June 1969, pp. 237–257.

SINGPURWALLA, N. D. and C. M. HARRIS, "On Estimation in Weibull Distributions with Random Scale Parameters," Vol. 16, No. 3, Sept. 1969, pp. 405–410.

SITGREAVES, R., S. HABER and H. SOLOMON, "A Demand Prediction Technique for Items in Military Inventory Systems," Vol. 16, No. 3, Sept. 1969, pp. 297–308.

SOLAND, R. M. and D. GROSS, "A Branch and Bound Algorithm for Allocation Problems in Which Constraint Coefficients Depend upon Decision Variables," Vol. 16, No. 2, June 1969, pp. 157–174.

SOLOMON, H., S. HABER and R. SITGREAVES, "A Demand Prediction Technique for Items in Military Inventory Systems," Vol. 16, No. 3, Sept. 1969, pp. 297–308.

SUURBALLE, J. W. and K. T. MARSHALL, "A Note on Cycling in the Simplex Method," Vol. 16, No. 1, Mar. 1969, pp. 121–137.

SWINSON, G. E. and P. H. RANDOLPH, "The Discrete Max-Min Problem," Vol. 16, No. 3, Sept. 1969, pp. 309–314.

THOMPSON, G. L. and O. MORGENSTERN, "An Open Expanding Economy Model," Vol. 16, No. 4, Dec. 1969, pp. 443–457.

VACHANI, M., "Determining Optimum Reject Allowances for Deteriorating Production Systems," Vol. 16, No. 3, Sept. 1969, pp. 275–286.

WEBB, S., "Interactions between the Experiment Designer and the Computer," Vol. 16, No. 3, Sept. 1969, pp. 423–433.

WOLLMER, R. D. and J. L. MIDLER, "Stochastic Programming Models for Scheduling Airlift Operations," Vol. 16, No. 3, Sept. 1969, pp. 315–330.

YUAN, J. S. C. and S. A. LIPPMAN, "Discounted Production Scheduling and Employment Smoothing," Vol. 16, No. 1, Mar. 1969, pp. 93–110.

* * *

☆ U.S. GOVERNMENT PRINTING OFFICE : 1970 OL—364-309

INFORMATION FOR CONTRIBUTORS

The NAVAL RESEARCH LOGISTICS QUARTERLY is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Manuscripts and other items for publication should be sent to The Managing Editor, NAVAL RESEARCH LOGISTICS QUARTERLY, Office of Naval Research, Washington, D.C. 20360. Each manuscript which is considered to be suitable material for the QUARTERLY is sent to one or more referees.

Manuscripts submitted for publication should be typewritten, double-spaced, and the author should retain a copy. Refereeing may be expedited if an extra copy of the manuscript is submitted with the original.

A short abstract (not over 400 words) should accompany each manuscript. This will appear at the head of the published paper in the QUARTERLY.

There is no authorization for compensation to authors for papers which have been accepted for publication. Authors will receive 250 reprints of their published papers.

Readers are invited to submit to the Managing Editor items of general interest in the field of logistics, for possible publication in the NEWS AND MEMORANDA or NOTES sections of the QUARTERLY.

NAVAL RESEARCH
LOGISTICS
QUARTERLY

DECEMBER 1
VOL. 16, NO.
NAVSO P-1278

CONTENTS

OFFICE OF NAVAL RESEARCH
Washington, D.C. 20360